University of Alberta

Capillary Electrophoresis for DNA Sequencing and Cytosine Methylation Analysis

by

Karl O. Voss

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Chemistry

Edmonton, Alberta

Spring 1998

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-29120-0

Canada

UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Capillary Electrophoresis for DNA Sequencing and Cytosine Methylation Analysis** submitted by Karl O. Voss in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dr. N. J. Dovichi

Dr. F. F. Cantwell

Dr. R. E. D. McClung

Dr. B. G. Kratochvil

Dr. A. R. Morgan

Dr. G. Drouin

December 19, 1997

In memory of Daryl Voss, whose life as a scientist, although cut far to short, sparked my own journey into science and technology.

ABSTRACT

DNA sequencing by capillary electrophoresis has two primary advantages over slab gel sequencing technology. More capillaries can be arranged in a single detection system than the number of lanes that can be used on a slab gel. Capillaries, because they are small and flexible, can be matched to nearly any sample vial format, allowing for fully automated operation of a CE instrument. These advantages can only be realized if sequencing separations done with capillary electrophoresis provide read-lengths equivalent to slab gels and provide them reproducibly.

I have identified two components of the capillary electrophoresis sequencing system that are responsible for band broadening and irreproducible sequencing separations. Replacement of the viscous sieving matrix between runs causes destruction of the capillary coating. Failure of the coating leads to increased band broadening and reduced read-lengths. Sequencing separations are performed at elevated temperatures to help reduce compressions that cause errors in sequencing data. However, small fluctuations in the temperature during the separations also lead to band broadening and reduced read lengths. Fluctuations in the capillary temperature must be less than 0.1 °C to avoid band broadening effects.

5-methylcytosine is found in the genomes of humans and other mammals where it is often referred to as the fifth base. DNA methylation is known to be involved in many important areas of biology but is poorly understood primarily because the positions of 5-methylcytosine in DNA sequences are difficult to determine. Sanger dideoxy sequencing chemistry is unable to differentiate between cytosine and 5-methylcytosine. Positions of methylcytosine can be determined by cytosine deamination-PCR (CD-PCR) chemistry. I

have developed a method to analyze CD-PCR products by direct cycle-sequencing and capillary electrophoresis. I have also shown that the PCR reaction used in CD-PCR chemistry preferentially amplifies the unmethylated DNA sequence, leading to errors in measured methylation levels. I have also shown that the PCR bias can be reduced by the addition of betaine to the CD-PCR chemistry. Finally, I have developed a CD-PCR system for methylation analysis of the p16 tumor suppressor gene.

Acknowledgments

I would like to express my appreciation to Dr. Dovichi for his guidance throughout my graduate program. I feel extremely fortunate to have had the opportunity to spend my time as a graduate student in his research group. I would also like to thank all the members of Dr. Dovichi's research group. In particular, Pieter Roos and Dawn Richards helped me immensely with my research. Kym Schriener also deserves special mention for much insight and the proofreading she provided for this thesis. Most importantly, I would like to express my gratitude to my family and my friends. They have made my four years as a graduate student a great learning experience as well an extraordinarily enjoyable period of my life. Thank you all.

Table of Contents

Chapter 1: Introduction 1

1.1 Introduction 2

1.2.1 Gel Electrophoresis 3

1.2.2 Mechanisms of DNA separation in polyacrylamide gels 3

1.2.3 Diffusion and band broadening in DNA sequencing 7

1.2.3 DNA sequencing by capillary electrophoresis 9

1.2.4 Capillary gel electrophoresis 9

1.2.5 Entangled polymers for DNA sequencing 10

1.2.6 Irreproducibility and the development CE-DNA sequencing 12

1.2.6.1 Sequencing sample composition and injection 12

1.2.6.2 Stability of polyacrylamide gels and sieving matrices ... 13

1.2.6.3 Stability of capillary coatings 13

1.2.6.4 Stability of the separation temperature 14

1.2.7 Reproducible DNA sequencing with replaced sieving matrix 17

1.3 DNA methylation analysis 18

1.3.1 Biological significance of 5-methylcytosine 18

1.3.1.1 (Cytosine-5)-Methyltransferase 18

1.3.1.2 Cytosine methylation and DNA structure 20

1.3.1.3 DNA methylation and gene control 22

1.3.1.4 CpG islands 22

1.3.1.5 Genomic imprinting and X-inactivation 24

1.3.1.6 DNA methylation, cancer and aging 25

1.3.2 Analysis of 5-methylcytosine in genomic DNA 26

1.3.2.1 Partial methylation and quantitative analysis of methylation status 26

1.3.2.2 Restriction endonucleases 27

1.3.2.3 Chemical sequencing 29

1.3.2.4 Ligation mediated PCR 29

1.3.2.5 Cytosine Deamination - PCR based methylation analysis 33
1.4 Bibliography 39

Chapter 2: DNA sequencing with poly-N,N-dimethylacrylamide: Stability of the capillary coating 47
2.1 Introduction 48
2.2 Experimental 49
2.2.1 Preparation of poly-N,N-dimethylacrylamide 49
2.2.2 Preparation of T-terminated sequencing samples 51
2.2.3 Preparation of silane-coated capillaries (Hjerten coating) 51
2.2.4 Preparation of Grignard coated capillaries (Novotny coating) 51
2.2.5 Instrumentation 52
2.2.6 Sequencing separations 52
2.2.6.1 Experiment 1: Multiple runs on a single capillary 52
2.2.6.2 Experiment 2: Single runs on aged capillaries (Hjerten coating) 54
2.2.6.3 Experiment 3: Multiple runs on a Grignard-coated capillary 54
2.2.3 Data analysis 54
2.3 Results and Discussion 55
2.3.1 Suitability of poly-DMA sieving matrix 55
2.3.2 Single termination sequencing reactions 58
2.3.3 Failure of refilled capillaries 58
2.3.4 Stability of the Hjerten coating 67
2.3.3 Stability of the Grignard-based coating (Novotny coating) 74
2.3.4 Stability of the poly-DMA sieving matrix 74
2.3.5 Peak widths and diffusion constants 74
2.4 Conclusion 77
2.5 Bibliography 78

Chapter 3: Thermal control in capillary electrophoresis DNA sequencing 80
3.1 Introduction 81
3.2 Experimental 83
3.2.1 Solid-state temperature controller 83
3.2.2 Temperature measurement 85
3.2.3 Preparation of T-terminated sequencing samples 88
3.2.4 Sieving matrix 88
3.2.5 Sequencing separations at a constant temperature 88
3.2.6 Sequencing separations with an AC temperature fluctuation 89
3.2.7 Data analysis 89
3.3 Results and Discussion 89
3.2.1 PID controller theory 89
3.3.2 The effect of temperature fluctuations on sequencing separations . 93
3.4 Conclusion 104
3.5 Bibliography 106

Chapter 4: Direct analysis of CD-PCR products by cycle-sequencing 107
4.1 Introduction 108
4.2 Experimental 109
4.2.1 Preparation of methylated DNA standard 109
4.2.2 Sodium bisulfite deamination of methylated and unmethylated pUC 19 109
4.2.3 Polymerase chain reaction 110
4.2.4 Cycle sequencing of CD-PCR product standards 110
4.2.5 Separation of the sequencing fragments by CGE 111
4.2.6 Data Analysis 111
4.3 Results and Discussion 112
4.3.1 Creation of methylated standard 114
4.3.2 Accuracy and precision of single-termination Thermosequenase chemistry 114

4.3.3 Direct CD-PCR sequencing *vs.* cloning 119
4.3.4 Problems encountered with direct sequencing of CD-PCR products 121
4.4 Conclusion 123
4.5 Bibliography 125

Chapter 5: Combating PCR bias in bisulfite-based cytosine methylation analysis: Betaine-modified CD-PCR 127
5.1 Introduction 128
5.2 Experimental 130
5.2.1 Creation of DNA standards with known methylation level 130
5.2.2 Calibration curve for CD-PCR chemistry 130
5.3 Results and discussion 130
5.4 Conclusion 137
5.5 Bibliography 142

Chapter 6: Methylation analysis of the p16/CDKN2 gene by CD-PCR 144
6.1 Introduction 145
6.2 Experimental 146
6.2.1 Preparation of cell line DNA 146
6.2.2 Confirmation of the existence of p16 in these cells lines 147
6.2.3 Deamination of cell line DNA 147
6.2.4 Polymerase chain reaction 147
6.2.5 Cycle sequencing of CD-PCR products 148
6.2.5 Separation of the sequencing fragments 149
6.3 Results and discussion 149
6.4 Conclusion 150
6.5 Bibliography 154

Chapter 7 155

7.1 Bibliograpy 159

List of Figures

Figure 1.1 Schematic plot of reduced mobility versus inverse of DNA length 6

Figure 1.2 Empirical determination of the entanglement threshold 11

Figure 1.3 Chemistry of the Hjerten coating 15

Figure 1.4 Chemistry of the Grignard coating 16

Figure 1.5 Structure of 5-methylcytosine 19

Figure 1.6 CpG sites and the maintenance activity of (cytosine-5)-methyltransferase 21

Figure 1.7 Structures of T-A-T and C-G-C+ base triads found in triplex DNA 23

Figure 1.8 Restriction digestion - Southern blotting methylation analysis 28

Figure 1.9 Restriction digestion - PCR methylation analysis 30

Figure 1.10 Reaction of cytosine with hydrazine 31

Figure 1.11 Ligation mediated PCR based methylation analysis 32

Figure 1.12 Reaction of cytosine with bisulfite 34

Figure 1.13 Site specific methylation analysis based on sodium bisulfite deamination of unmethylated cytosines 35

Figure 1.14 Creation of a degenerate CD-PCR from genomic DNA that differs only in methylation state 37

Figure 2.1 Apparatus used for the polymerization of DMA 50

Figure 2.2 System used to replace the poly-DMA sieving polymer solution 53

Figure 2.3 Electropherogram of a T-termination sequencing reaction. First run 56

Figure 2.4 Electropherogram of a T-termination sequencing reaction. Second run 57

Figure 2.5 Normalized resolution versus fragment length for experiment 1 60

Figure 2.6 Number of theoretical plates versus fragment length for experiment 1 62

Figure 2.7 Migration time versus fragment length for experiment 1 63
Figure 2.8 Peak width at half maximum for experiment 1 65
Figure 2.9 Base spacing for the first run from experiment 1 66
Figure 2.10 Normalized resolutions versus fragment length for experiment 2 69
Figure 2.11 Number of theoretical plates versus fragment length for experiment 2 71
Figure 2.12 Migration times versus fragment length for experiment 2 72
Figure 2.13 Peak width at half maximum versus fragment length for experiment 2 73
Figure 2.14 Normalized resolution versus fragment length for run 1 and run 10 of experiment 3 76
Figure 3.1 The effect of temperature on compressions in DNA sequencing 82
Figure 3.2 Heating unit used for DNA sequencing at elevated temperatures 84
Figure 3.3 PI controller circuit 86
Figure 3.4 Temperature monitor circuit diagram 87
Figure 3.5 Characteristic temperature profiles for a proportional feedback controller with too much and too little gain 91
Figure 3.6 Temperature oscillations at different gains 92
Figure 3.7 Effect of integral action on temperature profiles for a proportional feedback controller with too much and too little gain 95
Figure 3.8 Resolution versus fragment length for sequencing separations with different temperature fluctuations 97
Figure 3.9 Read length versus temperature fluctuation 99
Figure 3.10 Number of theoretical plates versus fragment length for different temperature fluctuations 101
Figure 3.11 Migration time versus fragment length for different temperature fluctuations 102
Figure 3.12 Peak width at half maximum versus fragment length for different temperature fluctuations 103
Figure 3.13 Model for temperature fluctuation induced band-broadening 105

Figure 4.1 CD-PCR and direct cycle sequencing chemistry for methylation analysis 113
Figure 4.2 Photograph of agarose gel of HpaII digest of unmethylated pUC 19 and methylated pUC 19 115
Figure 4.3 Electropherograms from A-termination sequencing of standard CD-PCR product mixtures 116
Figure 4.4 A section of an electropherogram of T-termination sequencing fragments created with Taq polymerase and Thermosequenase 120
Figure 4.5 Sequencing electropherograms of CD-PCR product from unmethylated DNA using Thermosequenase with 7-deaza-G and Thermosequenase with standard deoxyguanosine triphosphate . . . 122
Figure 4.6 Electropherograms of sequencing reactions of mixtures of CD-PCR products 124
Figure 5.1 Measured percentage of methylated DNA in standard by CD-PCR versus actual percentage of methylated DNA 132
Figure 5.2 Plot of observed mole fraction of methylated DNA vs. initial mole fraction of methylated DNA for conventional CD-PCR 134
Figure 5.3 Structure of betaine 139
Figure 5.4 Plot of observed mole fraction of methylated DNA vs. initial mole fraction of methylated DNA for betaine modified CD-PCR 141
Figure 6.1 T-termination sequencing electropherograms of CD-PCR products from A431 and HT29 cell lines 152
Figure 6.2 Photograph of agarose gel of PCR product from deaminated DNA using only the outer primer set 153

List of Tables

Table 2.1 Read lengths from the capillary refilling experiment 61

Table 2.2 Read lengths from experiment 2 70

Table 2.3 Read lengths from experiment 75

Table 3.1 Temperature fluctuations and characteristic frequency for differing PI controller gain 94

Table 3.2 Read lengths at different temperature fluctuations 98

Table 4.1 Percentage methylated DNA obtained by direct sequencing of CD-PCR product mixtures 117

Table 5.1 Calibration curve generated with conventional CD-PCR 131

Table 5.2 Calibration curve generated with betaine modified CD-PCR 140

List of Symbols and Abbreviations

μ^*	reduced electrophoretic mobility
μ	electrophoretic mobility
μ_o	electrophoretic mobility in free solution
μ_∞	limiting electrophoretic mobility
f_v	fraction of gel available to DNA molecules
M_D	molecular size of a DNA molecule in bases
M_a	molecular size of DNA molecule that equals average gel pore size
M	DNA molecular size beyond which mobility is size independent
C	gel concentration
E	electric field strength
$W_{1/2}$	peak width at half maximum
W	peak width at base
t	migration time
σ_t	standard deviation of peak
σ_{inj}	peak standard deviation due to injection effects
σ_{detc}	peak standard deviation due to detector effects
σ_D	peak standard deviation due to diffusion
σ_J	peak standard deviation due to Joule heating
σ_{other}	peak standard deviation due to unidentified components
D	diffusion constant
c^*	entanglement threshold for polymer solutions
$[\eta]$	intrinsic viscosity
η	viscosity
η_o	solvent viscosity

CE	capillary electrophoresis
EOF	electroosmotic flow
DNA	deoxyribonucleic acid
APS	ammonium persulfate
TEMED	tetraethylmethyethylenediamine
5-Me-C	5-methylcytosine
CDKN2	cyclin-dependent kinase inhibtor 2
PCR	polymerase chain reaction
LM-PCR	liga[redacted] mediated PCR
CD-PCR	cytosine deamination-PCR
DMA	dimethylacrylamide
PMT	photomultiplier tube
THF	tetrahydrofuran
HPLC	high performance liquid chromatography
TBE	tris-borate-EDTA buffer
EDTA	ethylenediaminetetraacetic acid
A	adenine
C	cytosine
T	thymine
G	guanine
GRIN	gradient index
PID	proportional-integral-derivative
ABI	Applied Biosystems
TMA	tetramethyammonium
TEA	tetraethylammonium
HLA	human lymphocyte antigens
TE	tris-EDTA buffer
PBS	phosphate buffered saline

NTP	nucleotide triphosphate

CHAPTER 1: INTRODUCTION

1.1 Introduction

The introduction of Sanger DNA sequencing chemistry[1] in 1977 is truly one of mankinds' great technological advances. It allows us to decode the information which actually makes us who and what we are, and record it on paper. Indeed, it is a great triumph of technology to decode the blueprint of life and store it in a medium as old as civilization itself.

The human genome contains approximately 3 billion bases of DNA sequence. The goal of the Human Genome Project is to determine the entire genomic DNA sequence of one human. The Human Genome Project has dramatically increased the demand for DNA sequencing throughput. Having determined the DNA sequence of one anonymous human, research interests will evolve into studying how the differences in DNA sequences actually make people different from each other. The switch of focus from the determination of a single DNA sequence to finding differences in this DNA sequence across many people will require a large increase in the amount of DNA sequence analysis. The demand for DNA sequencing throughput is likely to increase enormously in the coming years.

Sanger chain-termination chemistry and polyacrylamide gel electrophoresis allow for the determination of up to 500 bases of DNA sequence in a single analysis. Sequencing the entire human genome with traditional manual DNA sequencing, which uses radioactively labeled DNA, autoradiography, and the human eye, is not feasible. In 1986 Smith and coworkers added fluorescent labels to DNA sequencing primers[2]; this did not extend the amount of DNA that could be sequenced in a single analysis but did allow for automated detection and sequence determination. Another variation of fluorescent based DNA sequencing was introduced by Prober and co-workers, who added fluorescent labels to the chain terminating dideoxynucleotides[3]. The introduction of fluorescent based DNA sequencing by these two groups remains the biggest advance in DNA sequencing technology since Sanger's work 10 years earlier.

Slab gel electrophoresis with Smith and Prober's dye labeling remain the work-horse of the DNA sequencing world. The state of the art commercial DNA sequencing instrument is the Applied Biosystems model 377 slab gel sequencer. It can analyze 36

samples simultaneously, providing about 500 bases of sequence in 4 hours for each sample. Two or three runs a day are possible on this instrument, yielding a maximum throughput of 54000 bases per day. Data collection and sequence determination are automated with this instrument, however, new slab gels must be manufactured and installed manually between each run, and sequencing samples must be loaded onto the gel by a human operator. The total throughput of today's commercial DNA sequencing technology is reaching its limit. In order to meet the DNA sequencing demands of today and the future, new methodology and instrumentation must be developed.

1.2.1 Gel Electrophoresis

Sanger was fortunate to have the electrophoresis at his disposal to analyze the products of the chemistry that he had developed. Gel electrophoresis is a very powerful analytical tool. DNA sequencing separations are very demanding: DNA molecules hundreds of bases long must be resolved from those differing by only one base in length. The separations must be consistent enough to be analyzed by automated software without human input.

Electrophoresis, where molecules are caused to migrate through a solution by an electric field dates back to the 1930's when Tiselius constructed the first moving boundary electrophoresis aparatus[4,5]. Tiselius was awarded the Nobel prize for this work in 1948. Polyacrylamide gel electrophoresis, one of the key technologies that has made the revolutions in molecular biology possible, was introduced in 1959 when Raymond and Weintraub used it to separate proteins[6]. In 1961 Hjerten introduced agarose gel electrophoresis[7]. Polyacrylamide and agarose gel electrophoresis have become two of the most frequently used analytical separations technologies. For example, some large scale DNA sequencing laboratories analyze nearly 100,000 samples daily by these methods.

1.2.2 Mechanisms of DNA separation in polyacrylamide gels

It is the polyacrylamide gel which provides for the separation of DNA fragments of different size during electrophoresis. During electrophoresis, DNA molecules collide with strands of the polyacrylamide gel matrix, reducing their mobility. DNA molecules of different sizes interact differently with the gel matrix, allowing for a size based separation. There are a large number of reports that attempt to explain DNA separations

on a theoretical basis. Several reviews covering these reports are a place to start when reading the literature[8-12].

In 1958 Ogston developed a model that describes the spaces, or pores, that exist in a network of fiber shaped molecules[13]. Ogston's work forms the basis of the Ogston model for DNA separations in gels. This model assumes that the DNA molecules are spherical objects of a size equal to their radius of gyration and that the gel consists of a distribution of pore sizes between the gel molecules. Larger DNA molecules will only fit into a small fraction of the pores while smaller DNA molecules can fit into many more of the pores. The basis of the Ogston model is that the reduced electrophoretic mobility is proportional to the fractional volume available to the DNA molecule in the gel[14], according to the following equation:

$$\mu^* = \frac{\mu}{\mu_o} = f_V = \exp\left[-\frac{\pi(R+r)^2}{4aC^2}\right] \approx \exp\left(-\frac{M_D}{M_a}\right) \qquad (1.1)$$

where μ^* is the reduced electrophoretic mobility of the DNA fragment, μ is the actual electrophoretic mobility, μ_o is the mobility of the DNA fragment in free solution, f_V is the fraction of the gel that is available to the DNA molecule, R is the radius of gyration of the DNA molecule, r is the thickness of the gel strands, a is the average pore size of the gel, C is the gel concentration, M_D is the molecular size of the DNA molecule in bases, and M_a is the molecular size of a DNA molecule with a radius of gyration equal to the average pore size. The Ogston model assumes that the DNA molecule is spherical in shape and does not deform as it moves through the gel.

The Ogston model is valid only for small DNA molecules, with a radius of gyration smaller than the average pore size, and at low electric fields. Equation 1.1 predicts that the reduced mobility of DNA molecules decreases exponentially and eventually approaches zero as the molecular size increases. Experimental observations have shown that this is not the case[15-17]. Rather than having zero mobility, large DNA molecules have a limiting mobility for all DNA molecules larger than some threshold. In addition, rather than displaying an exponential dependence on size of the DNA,

mobilities of the intermediate sized DNA molecules vary as the inverse of the size of the DNA molecule. These observations are described mathematically as:

$$\mu^* \propto \frac{1}{M_D} \qquad M_a < M_D < M^* \tag{1.2}$$

$$\mu^* = \mu_\infty(E) \propto E \qquad M^* \ll M_D \tag{1.3}$$

where M^* is the size of the DNA fragment beyond which mobility becomes size independent, μ_∞ is the reduced mobility of DNA fragment of sizes $M_D >> M^*$ and E is the electric field strength.

These experimental observations led to the development of the biased reptation model[18, 19]. The biased reptation model assumes that the DNA molecules, being too big to move through the gel in a random coil conformation, move through the gel pores like a snake moves through grass. In the limits of small and very large DNA molecules the biased reptation model states:

$$\mu^* \propto \frac{1}{M_D} + \mu_\infty(E) \qquad M_a < M_D < M^* \tag{1.4}$$

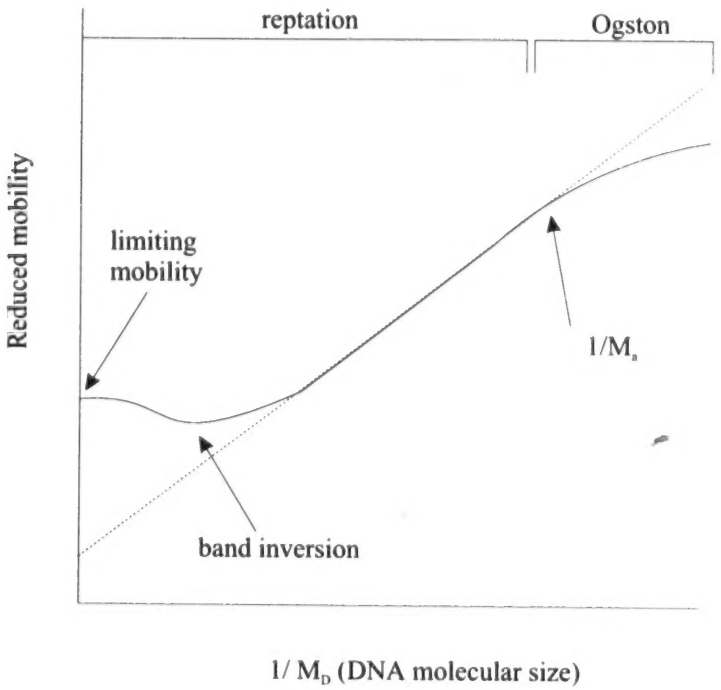
$$\mu^* \propto \mu_\infty(E) - \frac{1}{M_D} \qquad M_D \gg M^* \tag{1.5}$$

The limiting reduced mobility, μ_∞, predicted by the biased reptation model scales as:

$$\mu_\infty \propto E^2 \tag{1.6}$$

DNA mobility data can be plotted as reduced mobilities versus $1/M_D$ to clearly see the effects of the Ogston model and biased reptation model. A schematic of such a plot is shown in Figure 1.1. The regions of this plot that are described by the Ogston model and

Figure 1.1 Schematic plot of reduced mobility versus the inverse of the DNA fragment length. DNA lengths described by the Ogston and biased reptation regions are marked. Limiting mobility and band inversion region are also shown



the reptation model are marked. The linear region of the plot is for those intermediate sized DNA fragments described by equation 1.4. The point of departure from linearity as DNA fragments get smaller (marked $1/M_a$) is the point at which the size of DNA fragments equals the average pore size of the gel. Mobilities of fragments smaller than this size are described by the Ogston model. The curve approaches the limiting mobility (equation 1.6) as fragment size increases to infinity. The minimum of this curve is the point of band inversion. Band inversion, where the mobility of very large DNA fragments actually increases slightly, is a surprising prediction of reptation theory. The existence of band inversion has been confirmed experimentally[20].

Unfortunately, the prediction of the limiting mobility by reptation theory (equation 1.6) does not agree with experimental observation (equation 1.3). The biased reptation with fluctuations model was introduced by Viovy *et al.* to explain this discrepancy[16, 21]. This model assumes that the length of the reptating DNA molecule fluctuates as it passes through the gel. At low electric fields this model properly predicts the experimental limiting mobility.

Both the Ogston model and the biased reptation with fluctuations model are valid only for low field strengths. These models may not be valid at the high electric fields that are often used in capillary electrophoresis.

1.2.3 Diffusion and band broadening in DNA sequencing

Although there have been a many reports describing the mobilities of DNA in gels there are very few reports dealing with band broadening in DNA sequencing. The primary characteristic that defines a good DNA sequencing separation is its resolution. Resolution is considered sufficient for DNA sequencing if it exceeds 0.5 as calculated by:

$$Resolution = 2\times\frac{t_2 - t_1}{W_1 + W_2} \qquad (1.7)$$

where t_2 and t_1 are migration times for peaks from DNA fragments differing by one base in length and W_1 and W_2 are the full widths at the base of the same peaks. Resolution is reduced by decreased base spacing, as mobilities of the sequencing fragments approach their limiting mobility. Resolution is also reduced by broadening of the DNA bands as

they travel through the gel. Of these two factors, the latter is the more significant[22, 23].

The total variance of peaks in a sequencing separation can be expressed as:

$$\sigma_t^2 = \sigma_{inj}^2 + \sigma_{detc}^2 + \sigma_D^2 + \sigma_J^2 + \sigma^2_{other} \qquad (1.8)$$

where σ_t is the total variance in migration time of the DNA band, σ_{inj} is the variance due to sample injection, σ_{detc} is the variance due to the detector, σ_D is the variance due to diffusion of the DNA molecules and σ_J is variance due to thermal gradients caused by Joule heating. Except for σ_{detc}, the other components of peak variance are poorly understood. Band broadening is thought to be due primarily to diffusion, σ_D, and the temperature gradient across the gel due to Joule heating, σ_J. Peak variance due to diffusion is usually calculated with the Einstein relation:

$$\sigma_D^2 = 2Dt_m \qquad (1.9)$$

where D is the diffusion coefficient of the DNA molecule and t_m is the migration time. Calculations by Slater based on the biased reptation with fluctuations model indicate that the Einstein relation does not hold for DNA molecules in a gel under an electric field[23]. Because the DNA molecules orient with the electric field in order to pass through the gel, their diffusion constants in the longitudinal direction are higher than expected and are dependent on electric field strength. At high electric fields the diffusion constants are independent of the size of the DNA molecule. At the moderate electric fields used in slab gel sequencing the diffusion constants scale as:

$$D \sim \frac{E}{M_D^{1/2}} \qquad (1.18)$$

The fact that diffusion is more significant than suggested by the Einstein relation has

probably led to an overestimation of the importance of Joule heating[12].

1.2.4 DNA sequencing by capillary electrophoresis

One of the technologies that promises to greatly increase DNA sequencing throughput is capillary electrophoresis. Small diameter capillaries allow for efficient dissipation of the resistive heating (Joule heating) which occurs during electrophoresis. This fact allows much higher fields to be used in capillary electrophoresis than is possible with slab gels. Initially it was hoped that the use of high electric fields and capillary electrophoresis would allow sequencing separations to be performed in much less time than is required for slab gels. Unfortunately, the electric field that can be used in DNA sequencing is limited by the biased reptation effect as well as heat dissipation. Current CE-DNA sequencing technology allows for the determination of about 600 bases of DNA sequence in two hours[24]. This is only a two- to three-fold improvement over slab gel systems.

The major advantages capillaries have over slab gel systems stem from the size and flexibility of the capillaries themselves. Firstly, capillaries are small in diameter (< 200 μm), allowing many capillaries to be arranged in a single detection system. Designs exist for detectors capable of monitoring fluorescent emission from 864 capillaries simultaneously[25]. Secondly, capillary electrophoresis systems allow for automated injection of samples. The flexibility of capillaries makes possible designs of multicapillary instrumentation where injection is from a microtitre plate. In contrast, current slab gel systems can analyze, at most, 64 samples at a time. With a slab gel system, each sample must be loaded onto the gel individually by a human operator.

1.2.5 Capillary gel electrophoresis

DNA molecules have a similar charge-to-mass ratio regardless of their length. Therefore, DNA molecules of different lengths cannot be separated by classical free solution capillary electrophoresis[26]. CE separations of DNA can be accomplished by filling the capillary with a sieving matrix. The sieving matrix provides an extra frictional component that is proportional to the length of the DNA molecule, allowing for size based separations of DNA (Section 1.2.2). Capillaries filled with cross-linked polyacrylamide were first applied to separations of oligonucleotides in 1988[27]. The first

reports of DNA sequencing separations by CE were in 1990[28-31].

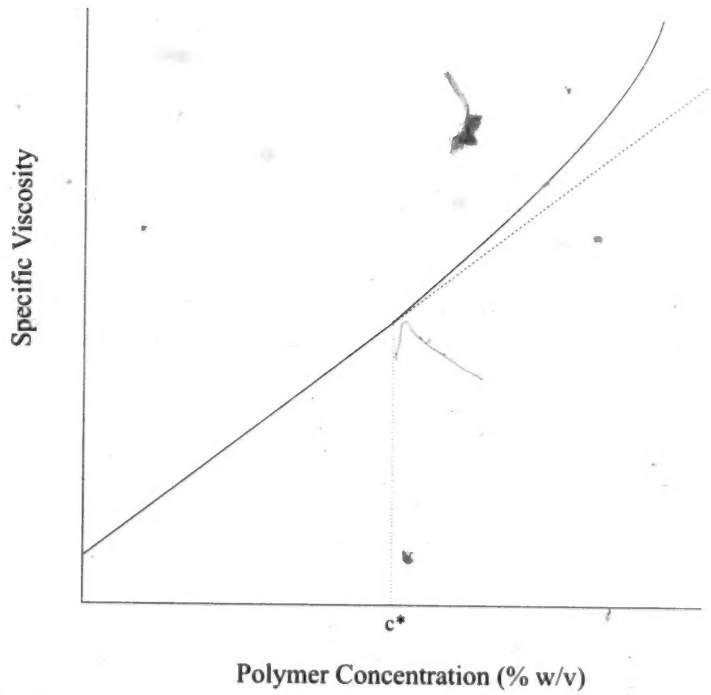
1.2.5 Entangled polymers for DNA sequencing

For large-scale DNA sequencing applications, crosslinked polyacrylamide gel sieving matrices are not ideal. Capillaries filled with polyacrylamide gel can be only used for a few runs after which the capillary must be replaced[32]. Fortunately, it is not necessary that the sieving matrix be a gel. Entangled polymer solutions can also function as sieving matrices for DNA sequencing. In 1977 Bode used non-crosslinked polyacrylamide for electrophoretic separations of proteins and nucleic acids[33, 34]. Crambach *et al.* also used non-crosslinked polyacrylamide for nucleic acid separations [35-37]. Non-crosslinked polyacrylamide has been used as a sieving matrix for CE-DNA sequencing[24, 38, 39]. Other polymer solutions such as polyethylene oxide[40] and modified polyethylene glycol[41] have also been used as a sieving matrix for DNA sequencing. Entangled polymer solutions have the potential to be pumped through the capillaries under high pressure[42], hopefully eliminating the need to replace the capillary after a few separations.

Gels, such as crosslinked polyacrylamide, have crosslinking bonds between the polymer chains. These bonds can be the strong, covalent type, as in polyacrylamide gels, or the weaker hydrogen bonds and hydrophobic interactions, as in agarose gels. Entangled polymer solutions do not have crosslinking bonds. They have what is sometimes referred to as topological crosslinks. At a polymer concentration above the entanglement threshold (c*) the polymer chains become entangled. Entangled polymer chains cannot move freely past each other; points where polymer chains are restricted in movement can be thought of as topological crosslinks. On a short time scale, the polymer solution has properties similar to a gel. The entanglement threshold can be measured empirically by plotting the specific viscosity versus concentration of the polymer solution. The point at which the plot deviates from linearity is approximately the entanglement threshold (Figure 1.2)[43]. Viovy and Duke claim that the entanglement is better determined from the intrinsic viscosity[44] according to the following equation[45]:

$$c^* \cong 1.5 \times [\eta]^{-1} \qquad (1.9)$$

Figure 1.2 Empirical determination of the entanglement threshold (c*) by plotting specific viscosity versus polymer concentration



where c^* is the entanglement threshold concentration and [η] is the intrinsic viscosity for the particular polymer. Intrinsic viscosity is defined as:

$$[\eta]=\lim_{c\to 0}\frac{1}{c}\left(\frac{\eta-\eta_o}{\eta_o}\right) \tag{1.10}$$

where c is the polymer concentration, η is the polymer solution viscosity, and η_o is the viscosity of the solvent. The intrinsic viscosity of a polymer solution can be determined experimentally by measuring the viscosity at several dilute polymer concentrations and extrapolating back to a zero polymer concentration.

1.2.7 Irreproducibility and the development CE-DNA sequencing

Considering the potential advantages of CE-DNA sequencing, it has not found widespread use. The acceptance of CE-DNA sequencing has been hindered by irreproducible separations. While occasional 1000 base reads obtained by CE have been reported[46] they remain difficult to reproduce. With a few exceptions, the CE-DNA sequencing literature has been primarily anecdotal. Only a few systematic studies have focused on CE-DNA sequencing separations in entangled polymers[47-50]. Unfortunately, these reports did not use a system where the polymer solutions were actually replaced and the capillary reused. Consistent read lengths of 500-600 bases by CE-DNA sequencing with reused capillaries and replaced sieving matrix have not been reported. Such studies have been extremely difficult to perform, mostly because of sheer lack of reproducibility.

There are four factors likely to play major roles in the reproducibility of the CE-DNA sequencing system: sequencing sample and injection reproducibility, stability and reproducibility of the entangled polymer sieving matrix, stability and reproducibility of the capillary coating, and stability of the separation temperature. The possible band-broadening effects that these components of the CE-DNA sequencing system have on sequencing separations has received almost no attention.

1.2.7.1 Sequencing sample composition and injection

The only systematic study of sample composition and injection in CE-DNA sequencing examines the effect of sample ionic strength on injection efficiency[49]. It does

not address band-broadening. In this thesis I do not address the effect of sample composition and injection on sequencing performance.

1.2.6.2 Stability of polyacrylamide gels and sieving matrices

As mentioned earlier, a replaceable sieving matrix is an essential component of any automated CE-DNA sequencing instrument. In addition to allowing for system automation, replaceable polymers can also eliminate the reproducibility problems that are likely to occur when polymerizing gels inside the capillary. However, the stability of the polymer solution is a concern. Righetti and co-workers have shown that polyacrylamide is susceptible to alkaline hydrolysis[51]. Righetti's group has developed two hydrophillic and hydrolytically stable monomers, N-acryloylaminoethoxyethanol[52, 53] and acryloylaminopropanol[54-56]. No studies describing the effect that hydrolysis of polyacrylamide has on CE-DNA sequencing separations have been reported.

In Chapter 2 I introduce poly-N,N-dimethylacrylamide as an alternative to polyacrylamide for CE-DNA sequencing. Poly-N,N-dimethylacrylamide is much less susceptible to hydrolysis than polyacrylamide, and is much easier to prepare than polymers made from Righetti's monomers.

1.2.6.3 Stability of capillary coatings

DNA sequencing requires that the capillary be coated to the reduce electroosmotic flow (EOF). Capillary coatings suppress EOF by derivatizating surface silanols and by establishing a zone of high viscosity near the capillary wall.

One main class of coatings is the dynamic coatings, where adsorption of some modifier, usually a polymer, on the capillary surface suppresses EOF. Dynamic coatings where polymers are added to the running buffer were first proposed by Hjerten[57]. Polymers such as methylcellulose[58] and polyvinylalcohol[59] have been used as dynamic coatings. A poly-N,N-dimethylacrylamide polymer with dynamic coating properties is available for DNA sequencing[60]. This polymer was used for the work in chapter 3. The efficacy of dynamic coatings is highly dependent on pH, which limits their usefulness for a wide variety of applications.

Permanent coatings rely on molecules that are chemically bonded to the capillary wall or immobilized films that cover the capillary walls. Films of polyvinylalcohol[61] and

cellulose[62] have been used as capillary coatings. These films are not covalently bonded to the capillary wall. Instead, they are semi-permanently attached to the capillary wall by baking polymer filled capillaries in an oven.

Covalently bonded, or permanent, coatings are the most commonly used method of coating capillaries for DNA sequencing experiments. Permanent coating have traditionally relied on silane chemistry to attach to the capillary wall. The silane agents used to coat the capillary often contain a functional group that allows them to be incorporated into an outer polymeric layer. Hjerten used this chemistry to create a polyacrylamide coated capillary[57]. The chemistry is based on attachment of the silanizing agent (γ-methacryloxypropyl)trimethyoxysilane to the capillary wall, followed by polymerization of an acrylamide solution inside the capillary (Figure 1.3). While this coating virtually eliminates EOF, it is not stable at high pH. Both the silane linkages and the polyacrylamide outer layer are susceptible to hydrolysis at an elevated pH. Hydrolysis of the coating results in an increase in EOF.

Preventing hydrolysis is a major goal of the capillary coatings research. Schmalzing *et al.* replaced the silane layer of the Hjerten coating with a cross-linked polysiloxane diol layer[63]. The polysiloxane layer contained vinyl functional groups which allowed it to be linked to a polyacrylamide outer layer. Novotny's group used a Grignard reaction to directly link vinyl groups to the capillary surface through silicon-carbon bonds[64](Figure 1.4).

Hydrolysis of the polyacrylamide outer layer can also result in an increase in EOF. Even if the linkage to the capillary surface is stable to hydrolysis, the polyacrylamide outer layer is susceptible to hydrolysis. However, the monomers developed by Righetti (Section 1.2.6.2) can be used instead of polyacrylamide to increase the stability of the coating.

There are no reports of systematic studies of capillary coating stability as the capillary is reused and the sieving matrix is replaced. In Chapter 2 I report the first study of the coating stability in refilled and reused capillaries for DNA sequencing.

1.2.7.4 Stability of the separation temperature

It is advantageous to perform DNA sequencing separations at 45°C to 55°C.

Figure 1.3 Chemistry of the Hjerten coating

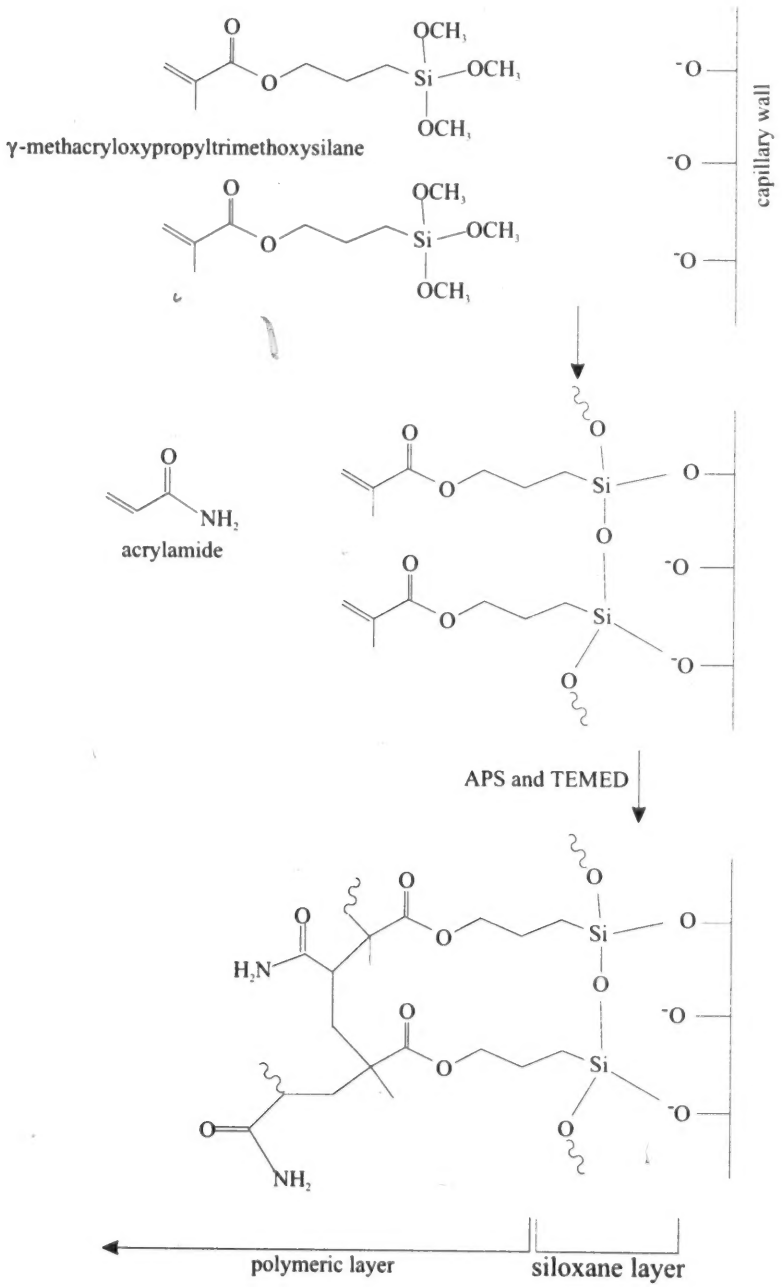
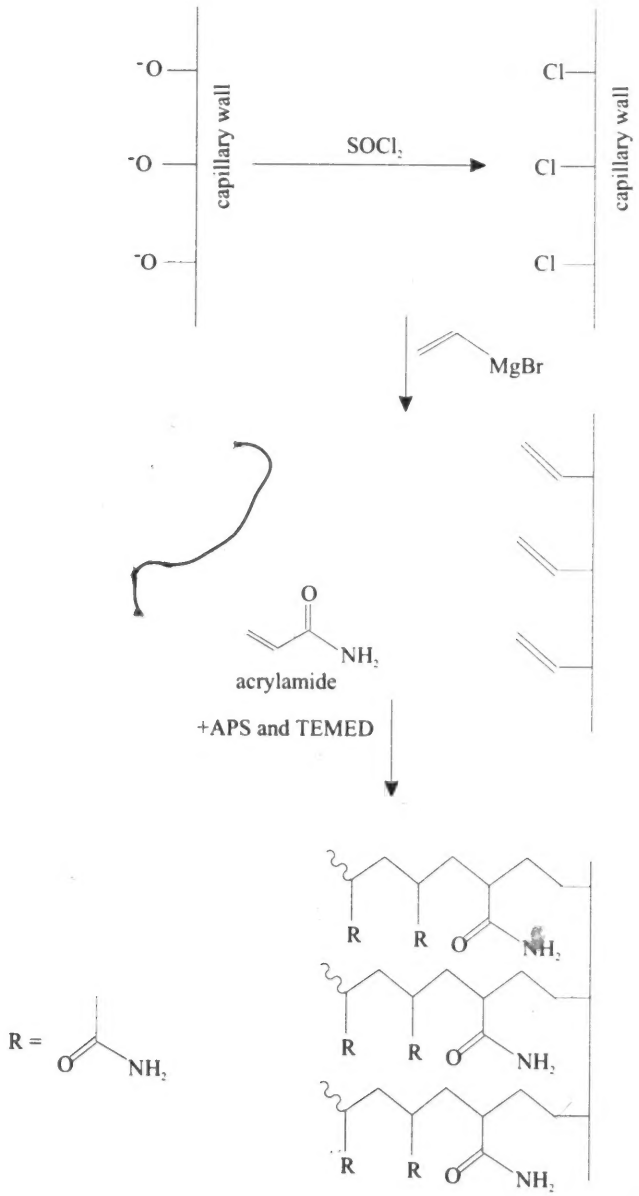


Figure 1.4 Chemistry of the Grignard coating



Elevated temperatures can eliminate some compressions in the CE-sequencing electropherograms by fully denaturing the sequencing fragments [65]. Performing CE-sequencing separations at elevated temperatures has the added benefit of increasing M^*, the length of DNA fragments after which the mobility becomes size independent (see Section 1.2.2) [66].

While it is generally acknowledged that CE-DNA sequencing separations should be performed at elevated temperatures no studies have attempted to determine what design specifications the heating unit must have. In particular, the precision with which the heating unit must control the temperature of the capillaries has not been determined. The experiment in Chapter 3 is the first report linking imprecise control of the capillary temperature to band-broadening and poor sequencing separations.

1.2.8 Reproducible DNA sequencing with replaced sieving matrix

The research described in chapter 2 and 3 is focused on obtaining reproducible separations with a replaced sieving matrix. Previous reports of CE-DNA sequencing used replaceable entangled polymer sieving matrices but did not perform multiple runs by replacing the polymer or did not achieve sufficient read lengths. These chapters also consider the identification and control of sources of band broadening that, if uncontrolled, lead to irreproducible separations.

1.3 DNA methylation analysis

Traditional DNA sequencing chemistry can determine the positions of four of the natural bases, A, T, G and C, in a DNA sequence. It cannot determine the position of a fifth natural base, 5-methylcytosine (5-Me-C). The structure of 5-Me-C is shown in Figure 1.5. It differs from standard cytosine only by the methyl group on the 5-position of the pyrimidine ring. Of the eukaryotes, the genomes of nearly all the vertebrates and plants as well as those of some lower organisms contain 5-methylcytosine.

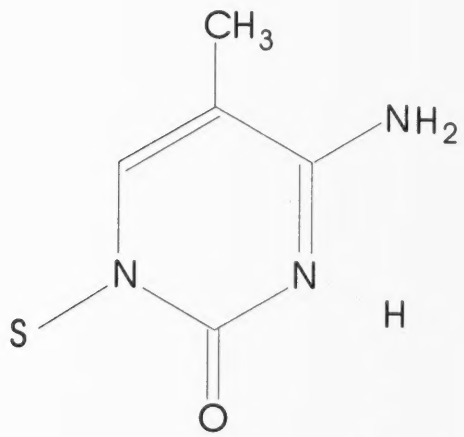
The study of DNA methylation is nearly 50 years old. The first report of the modified base 5-methylcytosine was by Hotchkiss in 1948[67], which was 5 years before the discovery of the double helical structure of DNA by Watson and Crick in 1953[68, 69]. Compared to the spectacular advances in molecular biology and genomics, the field of DNA methylation has not progressed much since then. Methylcytosine is a crucial component of the genetic material in the organisms in which it is found. Targeted homozygous deletions in the (cytosine-5)-methyltransferase gene, the enzyme responsible for creating 5-Me-C, prevents mouse embryos from developing past mid gestation[70]. However important it is, the actual function of 5-methylcytosine remains speculative. Patterns of DNA methylation are thought to be stable within a tissue but vary across tissue types. Methylation patterns change dramatically during embryogenesis, cellular differentiation and during carcinogenesis. While correlations exist between cytosine methylation and some biological processes, like gene regulation, the mechanism by which 5-methylcytosine is involved in these processes is unknown.

1.3.1 Biological significance of 5-methylcytosine

1.3.1.1 (Cytosine-5)-Methyltransferase

In mammalian genomes 5-methylcytosine is found primarily in CpG dinucleotides (CpG=5' CG 3'). Of the approximately 5×10^7 CpG dinucleotides in the human genome[71] approximately 60% are methylated[72]. Other methylation sites, such as CpNpG are reported in human DNA[73]. Cytosine methylation patterns in eukaryotic organisms are created and maintained by (cytosine-5)-methyltransferase. This methylase enzyme is responsible for the transfer of a methyl group from S-adenosylmethionine to the 5- position of the cytosine ring. Only a single methyltransferase has been detected in

Figure 1.5 Structure of 5-methylcytosine



S = deoxyribose

eukaryotes; it is assumed that the same enzyme is responsible for both maintenance and *de novo* methylation activities. Maintenance methylation occurs following DNA replication. CpG dinucleotides that are methylated in both strands of the DNA molecule are called fully methylated CpG's. Following DNA replication, a CpG site will be methylated on only one of the DNA strands. This is known as a hemi-methylated CpG site. Very soon after replication methyltransferase recognizes the hemi-methylated CpG and converts it into a fully methylated CpG[74]. This is known as the maintenance activity of the methyltransferase (Figure 1.6). The *de novo* mode of methyltransferase targets unmethylated CpG sites and converts them into hemi-methylated CpG's. The *de novo* activity of the enzyme is responsible for the development of tissue-specific methylation patterns during enbryogenesis. The methyltransferase maintenance activity is 100 fold greater than the *de novo* activity *in vitro*, but the relative activities *in vivo* are unknown[75].

The cDNA for the human methyltransferase has been cloned[76, 77]. It is a 5194 base transcript potentially encoding for a 1495 amino acid polypeptide of 169 kDa. Of the approximately 1500 amino acids in the human methyltransferase the C-terminal 500 amino acids are responsible for transferring the methyl group to cytosine[78]. The 1000 N-terminal amino acids, which contain two zinc binding domains and one DNA binding domain[79], imparts specificity of the enzyme for hemi-methylated substrates.

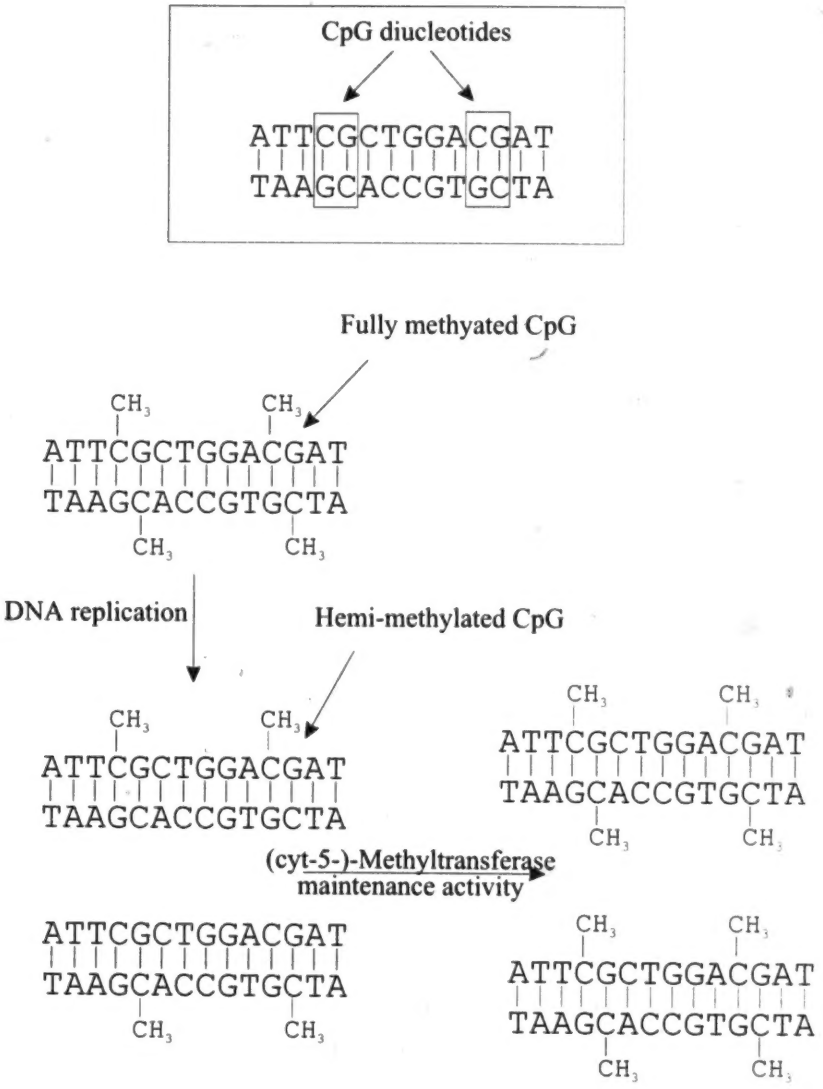
The enzymology of the methyltransferase is reviewed by Smith[75]. Currently little is known about the processes that regulate the activity of the enzyme. While some secondary structures in the target DNA molecules increase the rate of *de novo* methylation[75, 80] how the enzyme is targeted to *de novo* methylate specific DNA sequences during growth and development is unknown.

1.3.1.2 Cytosine methylation and DNA structure

5-methylcytosine is known to influence the secondary structure of DNA. It slows the rate of cruciform formation when methylcytosine is in the central loop region of the cruciform forming sequence[81]. Cruciforms are analogous to intermediate structures that may form during DNA recombination events. Methylation of cytosines can both increase and decrease curvature of DNA sequences[82, 83].

Methylcytosine also stabilizes DNA triple helices (triplex DNA). Triplexes are

Figure 1.6 CpG sites and the maintenance activity of (cytosine-5)-methyltransferase



DNA structures where three strands of DNA are hydrogen bonded together. Triplex with T-A-T and C-G-C+ base triads (shown in Figure 1.7) form only at pH lower than 6. Substituting 5-methylcytosine for cytosine in the third strand allows triplex formation at physiological pH[84].

The Watson-Crick double helix, otherwise known as B-Form DNA, has a right-handed twist to the helix. DNA composed of mainly alternating GC or AC dinucleotide repeats can form a double helix with a left-handed twist, known as Z-DNA, in solutions of high ionic strength and magnesium concentration[85]. However, DNA composed of alternating (5-Me-C)-G repeats can form Z-DNA structures at considerably lower ionic strength[86].

While all of the DNA structures discussed above exist *in vivo,* their function and significance is still speculative. One thing is certain, however: It is the three-dimensional structure of DNA, assembled into chromatin and associated with the all the DNA binding proteins that ultimately determines which genes are expressed and how the information encoded in DNA controls the cell machinery. Any event which influences the structure of DNA, such as DNA methylation, is likely to be of fundamental importance to the processes of life.

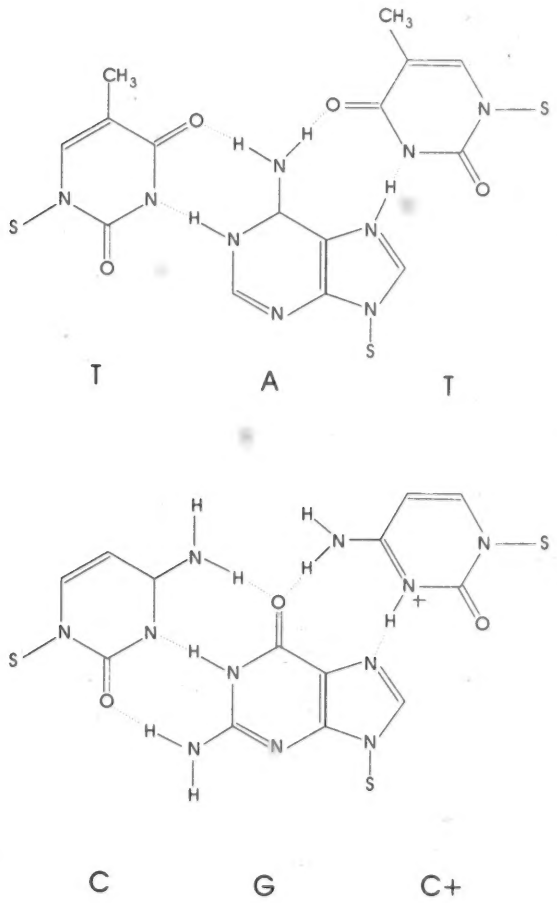
1.3.1.3 DNA methylation and gene control

Most interest in DNA methylation is due to evidence that methylation is involved in controlling gene expression. A consistent picture of DNA methylation patterns has developed through many studies of the methylation states of DNA sequences from various genes. Methylation patterns differ between different tissue types but are stable within tissue types. Methylation patterns change as cells differentiate, age, and undergo tumorgenesis.

1.3.1.4 CpG islands

The CpG dinucleotide is under-represented in the mammalian genome. Although many CpGs are found in repetitive satellite DNA sequences, which are usually methylated, CpG sequences are also clustered near the 5' end of genes. Such CpG clusters are known as CpG islands. CpG islands are characterized by a CpG/GpC ratio of about 1.0 and a G + C content of 55% to 70%, as compared to 0.2 and 40% respectively, for the

Figure 1.7 Structures of T-A-T and C-G-C+ base triads found in triplex DNA



bulk DNA[87]. They are approximately 1kb in length and contain the promotor regions of genes and binding sites for various transcription factors. Experiments have shown that expression of a gene is eliminated when the gene's CpG island is methylated[88]. Not all genes contain CpG islands but those autosomal genes that do, including all known housekeeping genes, have CpG islands that are unmethylated in normal tissue[88]. Most tissue-specific genes do not have CpG islands but do show some degree of correlation between the extent of methylation and transcriptional activity.

Studies suggest that methylation represses transcription either by interfering with the binding of transcription factors or by directing the assembly of the DNA into a closed chromatin structure. Methylation disrupts the DNA-binding activity of several transcription factors[89, 90]. The human nuclear proteins MDBP-1[91] and MeCP[92] have greatly enhanced binding to methylated DNA sequences. It is speculated that these proteins bind to methylated DNA regions, causing the DNA to assemble into condensed chromatin which is transcriptionally inactive[93, 94].

1.3.1.5 Genomic imprinting and X-inactivation

DNA methylation is an example of a system for maintaining and expressing epigenetic information, that is, genetic information extra to the primary DNA sequence. Genomic imprinting is one example of epigenetic inheritance, where an organism displays differential expression of the genes originating from its mother and its father. The most studied examples of imprinting in mammals is X-chromosome inactivation (reviewed by Monk[95]). X-inactivation occurs in female mammals, where some genes on the paternally acquired X chromosome are preferentially inactiviated. This ensures that both males and females receive equal genetic dosages of X-linked genes. The inactivation of housekeeping genes on the X-chromosome correlates with the methylation of their CpG islands. X-chromosome inactivation occurs in early embryogenesis, near the time of uterine implantation[96]. The inactivated genes on the X-chromosome remain inactive throughout all cells of the mammal, except for the germ-line cells. Genes on the inactive X chromosome are not expressed even though they are exposed to the same nuclear environment and transcription factors as the expressed genes on the active X chromosome. Exactly how the embryonic cells know to inactivate the paternal X-

chromosome is still a matter of speculation, as is the timing of the cytosine methylation relative to the actual silencing of the genes.

1.3.1.6 DNA methylation, cancer and aging

Recently there has been considerable interest in the possibility that aberrant methylation may play a role in carcinogenesis. Three possible mechanisms have been proposed to explain how methylation may be involved in the processes that transform a normal cell into a cancerous one. The first, and well accepted theory, is that 5-methylcytosine is a mutational hotspot. Spontaneous deamination of cytosine results in uracil, which is removed from the genome by uracil DNA glycosylase. In contrast, spontaneous deamination of 5-methylcytosine results in thymine, which remains in the genome and causes a C to T transition mutation[97]. For example, the most commonly mutated gene in human tumors is the p53 gene[98]. If mutations of the p53 gene were entirely random only 3% of mutations would be expected to occur at CpG sites. However, of more than 5000 p53 mutations found in many tumor types, 30% occur at CpG sites[99]. 5-Methylcytosine is a potent mutagen.

DNA methylation may be involved in carcinogenesis through its role in gene regulation. Hypomethylation of proto-oncogenes such as *raf, c-myc, c-fos, c-H-ras* and *c-K-ras* have been reported[100-102]. Presumably, oncogenes lacking proper methylation are more active than they should be, which could lead to cancer. Recently, a number of reports have suggested that methylation of the CpG islands of tumor suppressor genes can silence their transcription[103-110]. It is believed that the *de novo* activity of the methytransferase slowly methylates cytosines in the CpG islands of tumor suppressor genes. Once methylated, the high maintenance activity of the methyltransferase will ensure that the CpG site remains methylated in all subsequent daughter cells. Over time, the number of methylated sites may build up and eventually reduce or silence the transcription of that gene.

The evidence for a role of methylation of both tumor suppressor genes and oncogenes as a cause of cancer is correlative only; to date there is no direct evidence that cytosine methylation can cause cancer. This is not surprising considering that the mechanisms which describe the role methylation plays in gene control also remain

speculative. Chapter 6 describes a system for the analysis of the methylation state of the p16/CDKN2 tumor suppressor gene, one of the many genes where methylation is believed to play a role in carcinogenesis.

DNA methylation may also be involved in cellular aging[111, 112] and age related diseases. For example, differential methylation has been reported in the promotor region of the β-amyloid precursor protein gene in human brain tissue[113, 114]. This gene has been implicated in Alzheimers's disease

1.3.2 Analysis of 5-methylcytosine in genomic DNA

Despite the fact that DNA methylation is associated with very important areas of biology and was first studied nearly 50 years ago, surprisingly little is known about the function of DNA methylation. This is due, mostly, to the limited analytical tools available for determining methylation patterns.

It is currently believed that cytosine methylation acts as a gross genetic control mechanism, where genes are maintained in an inactive state by DNA methylation that induces or supports a condensed chromatin structure. This model is based almost entirely on data about the methylation patterns of genes which was acquired with analytical techniques that were able to provide the methylation state of only a few cytosines in the entire gene. Such an incomplete analysis of methylation state is inadequate considering that a single methylated cytosine can interfere with the binding of a transcription factor. Better analytical technologies are a needed to obtain a clearer understanding of DNA methylation.

1.3.2.1 Partial methylation and quantitative analysis of methylation status

Standard DNA sequencing chemistry determines the base sequence of a region of DNA. DNA sequencing requires that the population of DNA molecules used as a template for the sequencing reaction have an identical sequence. DNA sequencing is usually performed using cloned DNA as a template, which has a homogeneous sequence. Aside from infrequent mutations, the primary DNA sequence is identical in all cells of a single organism. Therefore, a single DNA sequence is adequate for the analysis of a particular area of DNA in a single organism.

In contrast, the positions of 5-methylcytosine in the genome of a single organism

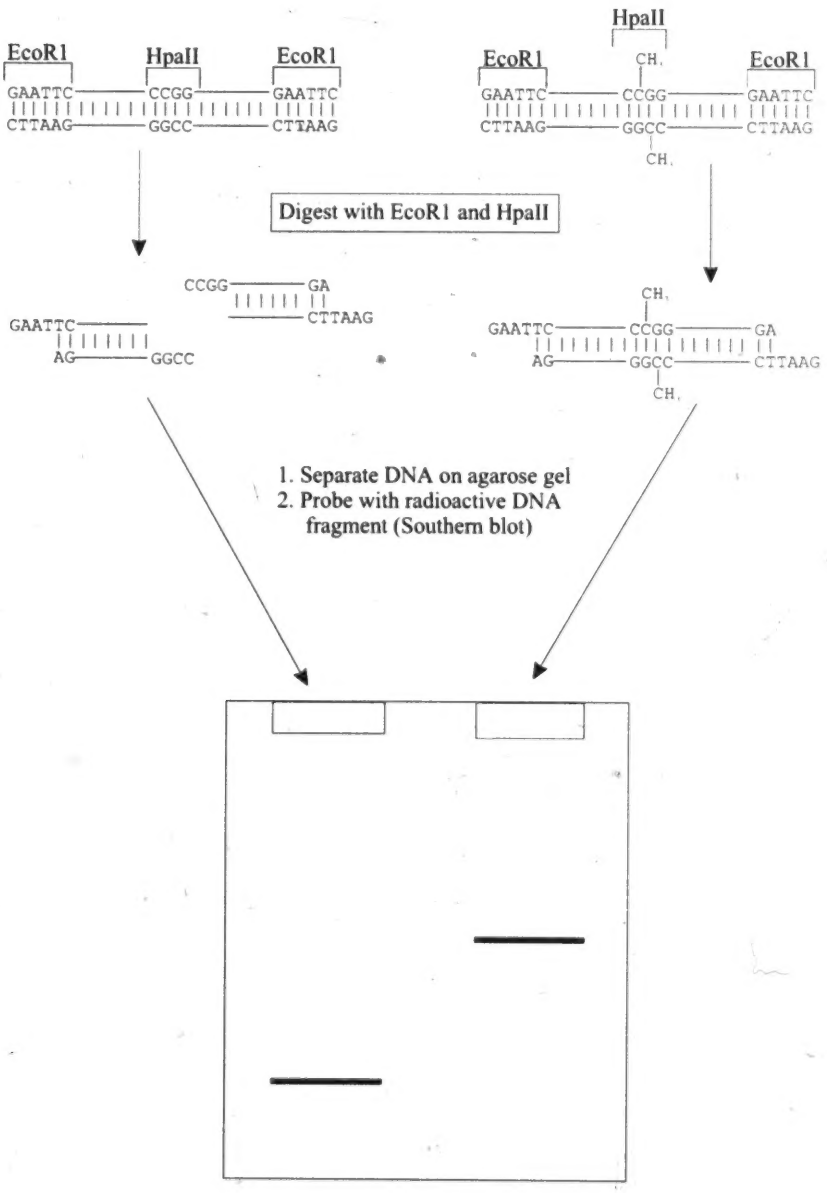
vary across tissues, and may vary within a population of cells of the same type. The methylation status of cells also changes as they undergo differentiation, oncogenesis and possibly ageing. Because of this, any analytical technique that determines the positions of 5-methylcytosine must either be capable of analyzing a single cell or must quantitatively determine the extent of methylation at each cytosine.

The term "partially methylated" is often used in the literature to refer to regions of DNA sequence where methylation has been detected at some cytosines but not at other cytosines. In this case "partially methylated" is used in a qualitative sense and does not refer to the extent of methylation at a particular cytosine. With respect to quantitative analysis of methylation levels, the term "partially methylated" is better suited to describe the methylation status of an individual cytosine. I describe an individual cytosine as partially methylated in a particular sample of cells if it is methylated in some of the cells but unmethylated in other cells. Analysis of partially methylated DNA samples requires an analytical technique that is capable of quantitative analysis of methylation levels. Such a technique must be able to determine the percentage of DNA which is methylated at each and every cytosine in a particular DNA sequence.

1.3.2.2 Restriction endonucleases

The classical techniques for the sequence specific analysis of 5-methylcytosine utilize methyl-sensitive restriction enzymes[115]. Some restriction enzymes cannot cleave double-stranded DNA when one or more cytosines in their recognition site is methylated. The restriction digestion-Southern blotting methylation analysis system is summarized in Figure 1.8. Genomic DNA of interest is digested with a methyl-sensitive restriction enzyme, such as *HpaII*, and appropriate methylation insensitive restriction enzymes, such as *EcoRI*. The digested DNA is separated on an agarose gel and hybridized to a radioactive probe complementary to the DNA sequence of interest. Although this analysis is fairly time-consuming, it is simple to perform and provides quantitative information about partially methylated cytosines. Unfortunately, the entire arsenal of available restriction enzymes is not able to analyze all the possible sequences that may contain a CpG dinucleotide. Further, restriction enzymes cannot distinguish between a fully methylated and hemi-methylated cytosine. The restriction digestion-Southern blotting

Figure 1.8 Restriction digestion - Southern blotting methylation analysis



agarose gel and Southern blot

methodology has poor sensitivity; several micrograms of genomic DNA are required for an analysis. A variation of the restriction analysis system (Figure 1.9), where the Southern blotting step is replaced with a PCR across the restriction site, has very high sensitivity[116]. Because of its sensitivity and short analysis times, this technique is the most commonly used method of methylation analysis in recent literature. However, the use of a PCR to probe for restriction fragments is qualitative only, and does not allow quantitation of partial methylation levels.

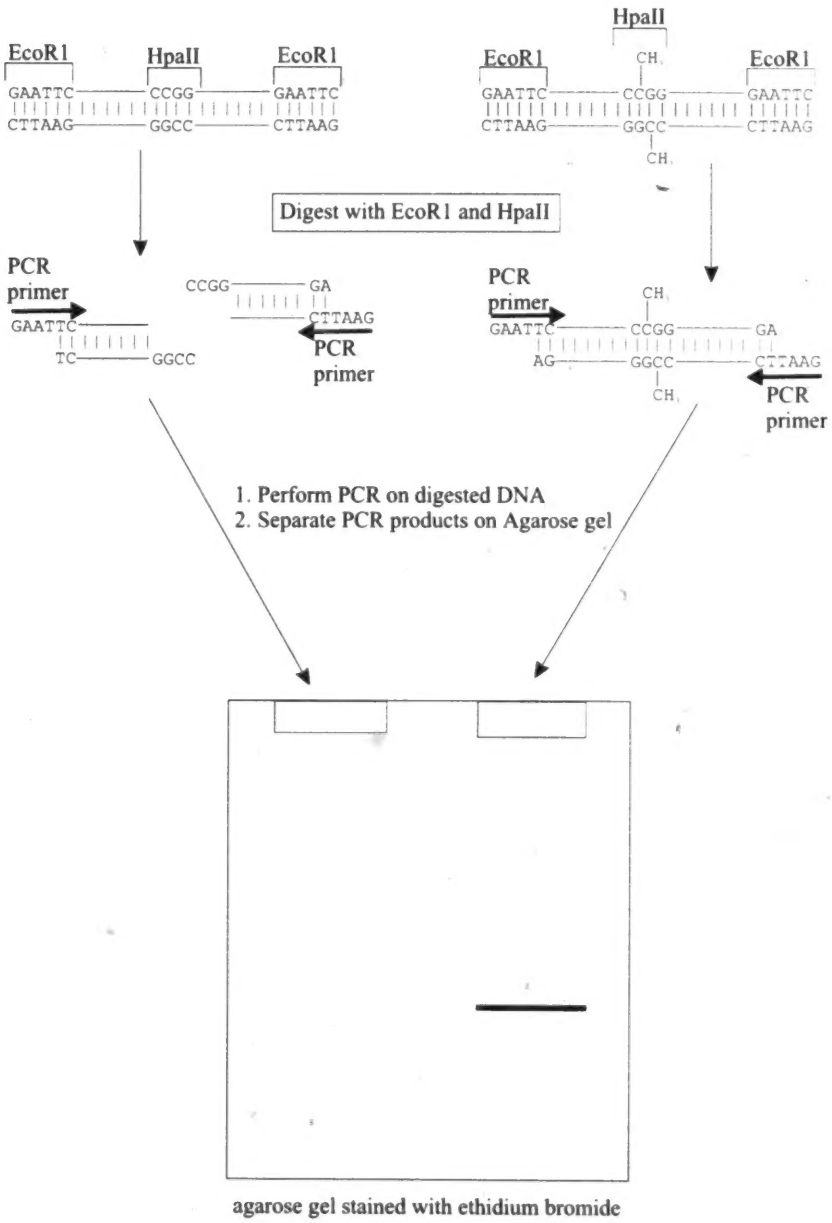
1.3.2.3 Chemical sequencing

Maxam-Gilbert style chemical DNA sequencing[117] uses hydrazine-alkali cleavage of the DNA backbone at cytosine residues (Figure 1.10). Hydrazine is much less reactive with 5-methylcytosine residues. A 5-methylcytosine will leave a gap in the C lane of a Maxam-Gilbert sequencing ladder. This method is only useful for sequencing highly purified DNA from very small genomes. To allow for analysis of any DNA sample, Church and Gilbert used labeled probes to visualize only the sequencing fragments of interest, and not the background DNA, after separation by electrophoresis[118]. To improve the sensitivity of chemical sequencing, Saluz and Jost followed the chemical cleavage step with a linear amplification by Taq polymerase[119]. This step enhances the signal in the same as it does in cycle sequencing. Although chemical sequencing may provide quantitative information about partially methylated sites, its use has been rarely reported, due to its poor sensitivity and use of toxic hydrazine.

1.3.2.4 Ligation mediated PCR

Site-specific determination of methylated cytosine by ligation-mediated PCR (LM-PCR)[120, 121] uses PCR amplification of hydrazine-treated DNA to obtain high sensitivity, site-specific methylation analysis. This process, detailed in Figure 1.11, involves hydrazine cleavage at unmethylated cytosines, primer extension to create blunt ends, and ligation of a double-stranded linker to the newly formed blunt end. These fragments, now containing the sequence of the linker at one end and a known sequence at the other end, are used as a template for a PCR reaction. Primers chosen for the PCR reaction match the double-stranded linker and the primer used to create the blunt-ended DNA. The PCR products are analyzed on a denaturing polyacrylamide gel.

Figure 1.9 Restriction digestion - PCR methylation analysis



agarose gel stained with ethidium bromide

Figure 1.10 Reaction of cytosine with hydrazine

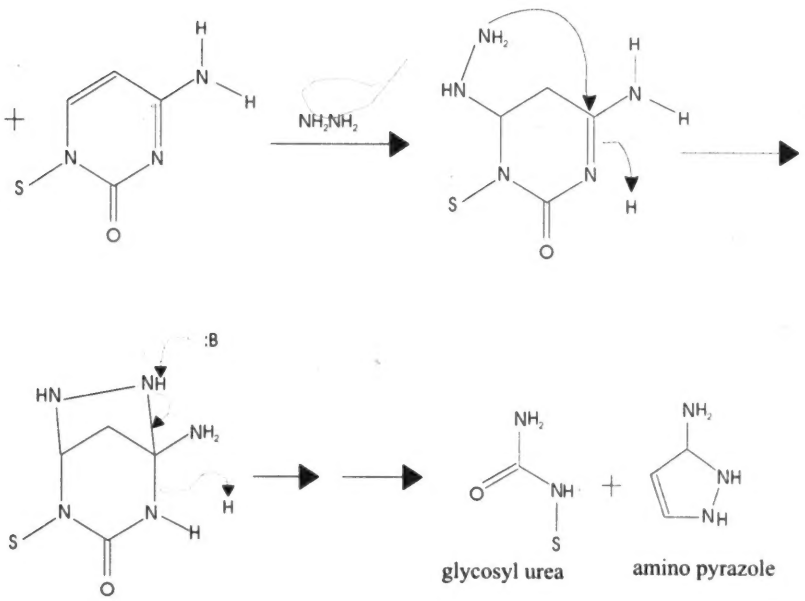
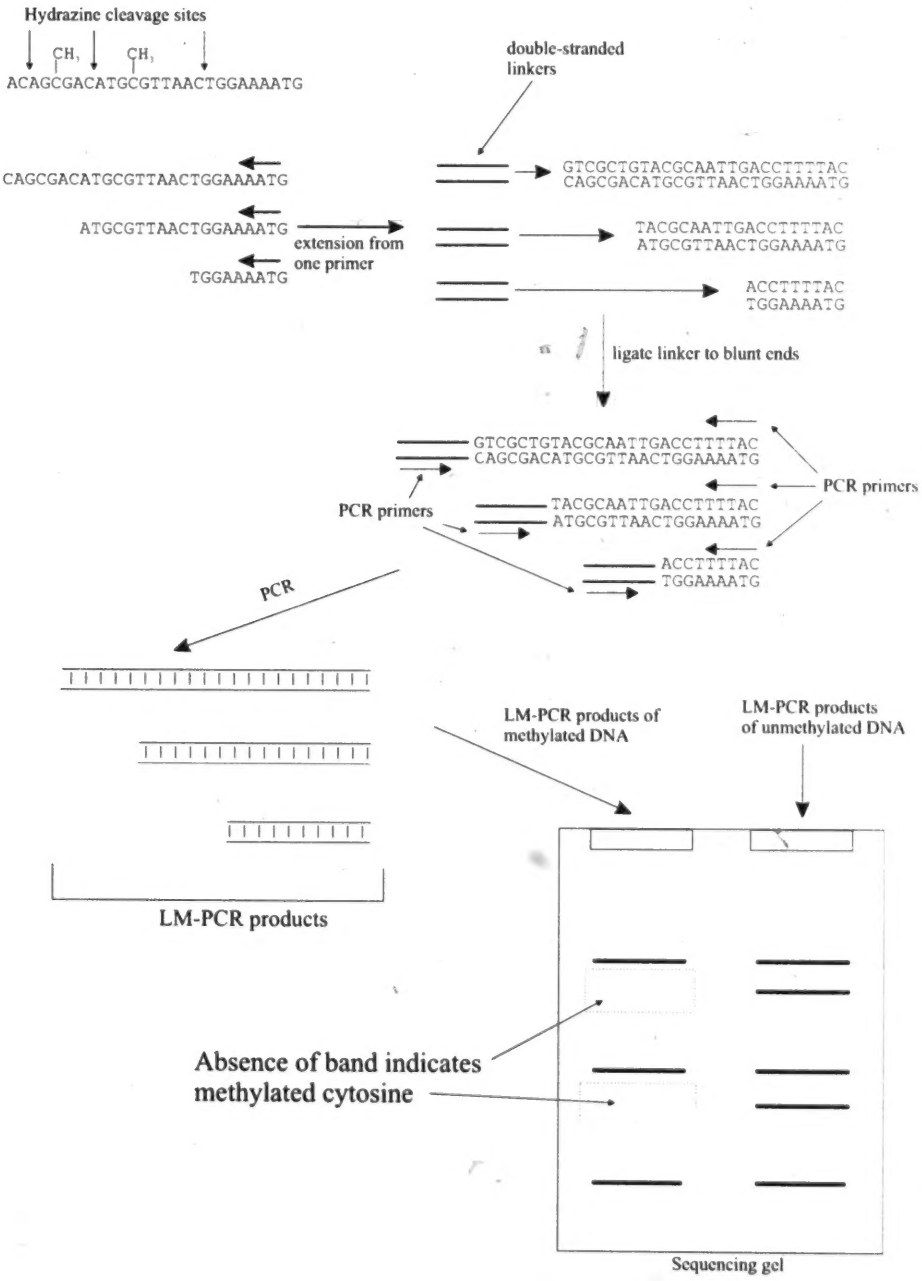


Figure 1.11 Ligation mediated PCR based methyation analysis



Because a molecule of template is consumed each time a hydrazine cleaves at a cytosine, several thousand molecules of genomic DNA are required to obtain statistically significant results. This defines the theoretical limit of sensitivity for LM-PCR. Ligation-mediated PCR, which relies on PCR to simultaneously amplify many different DNA fragments, is not a quantitative technique. Ligation-mediated PCR is rarely used for methylation analysis because it was quickly superseded by cytosine deamination-PCR based methylation analysis method.

1.3.2.5 Cytosine Deamination - PCR based methylation analysis

The cytosine deamination-PCR (CD-PCR) based methylation analysis technology has become the method of choice to determine the methylation state of every cytosine in a DNA sequence. Cytosine and 5-methylcytosine have different reactivities with sodium bisulfite. The reactions of sodium bisulfite with uracil, cytosine and their analogs was reported in 1970 by Hayatsu[122]. At high concentrations of bisulfite (~4 M) and low pH (~5) bisulfite adds quantitatively across the 5-6 double-bond of cytosine residues in single-stranded DNA[123]. Under the same conditions, less than 2% of the 5-methylcytosine residues react with the bisulfite. Bisulfite addition is a reversible reaction but it renders cytosine (and 5-methylcytosine) susceptible to hydrolytic deamination; incubation for 16 to 24 hours at a temperature of 50 °C allows for the deamination of all cytosines in a single-stranded DNA molecule. Deamination is an irreversible reaction; subsequent base-catalyzed elimination of bisulfite results in uracil (or thymine, in the case of the few 5-methylcytosines that reacted). The chemistry of this process[124] is shown in more detail in Figure 1.12.

Selective deamination with sodium bisulfite provides the basis for site-specific methylation analysis as shown in Figure 1.13. Genomic DNA is denatured, and subjected to deamination by sodium bisulfite as described above. Cytosines are converted to uracil by deamination. 5-methylcytosine remains as 5-methylcytosine. The deaminated DNA, which is no longer complementary, is used as a template for a PCR reaction. Primers for the PCR are chosen to amplify a DNA sequence of interest from one of the strands of the deaminated genomic DNA. To amplify both the methylated and unmethylated DNA sequence, primers must be chosen that will avoid CpG's in the DNA sequence. These

Figure 1.12 Reaction of cytosine with bisulfite

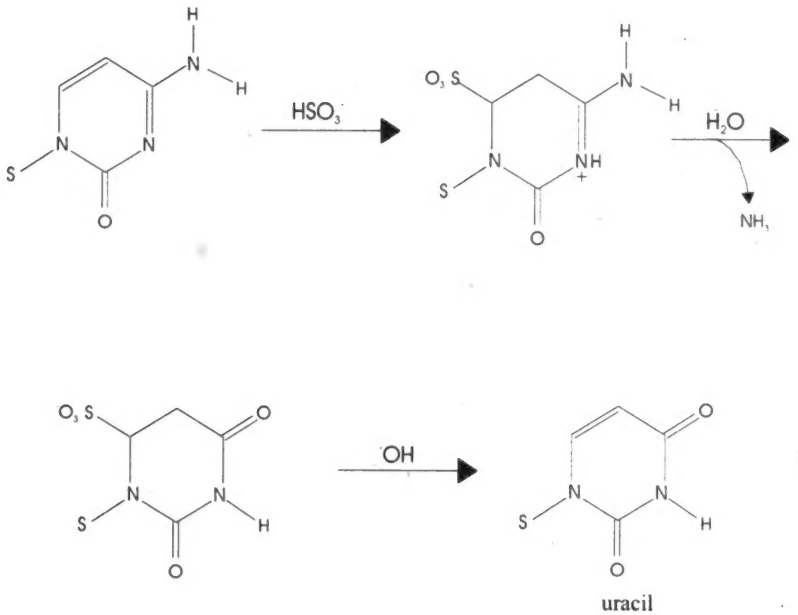
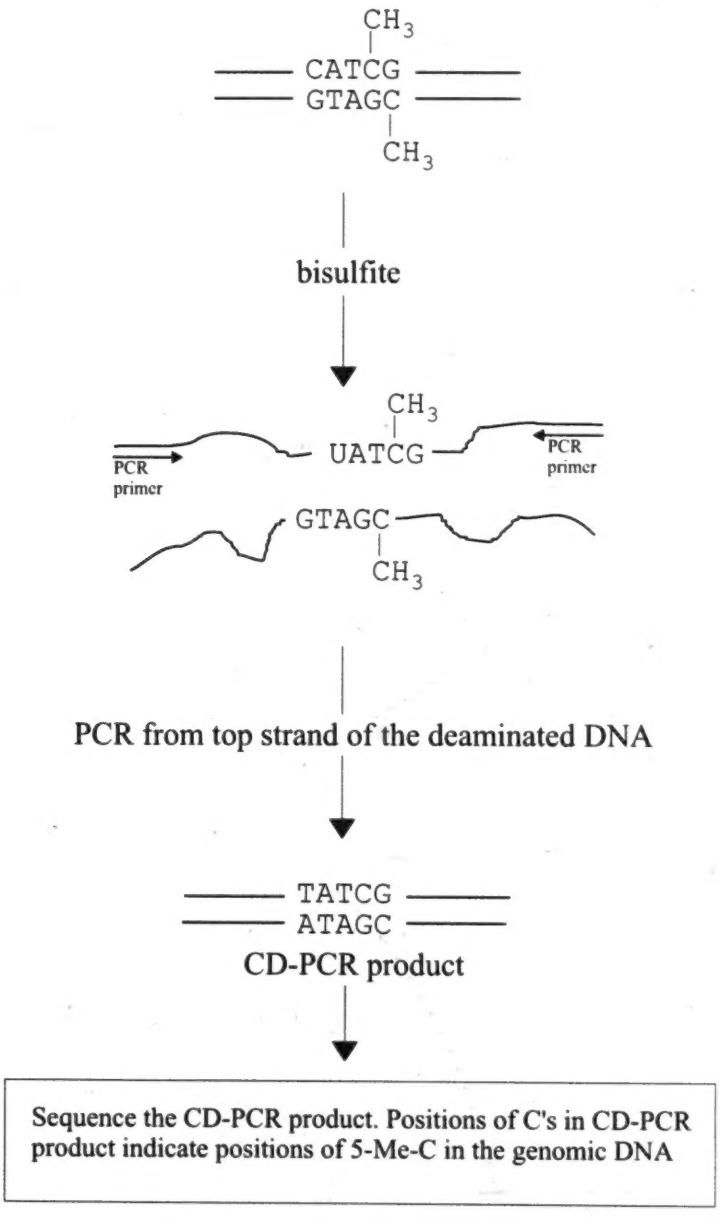


Figure 1.13 Site-specific methylation analysis based on sodium bisulfite deamination of unmethylated cytosines



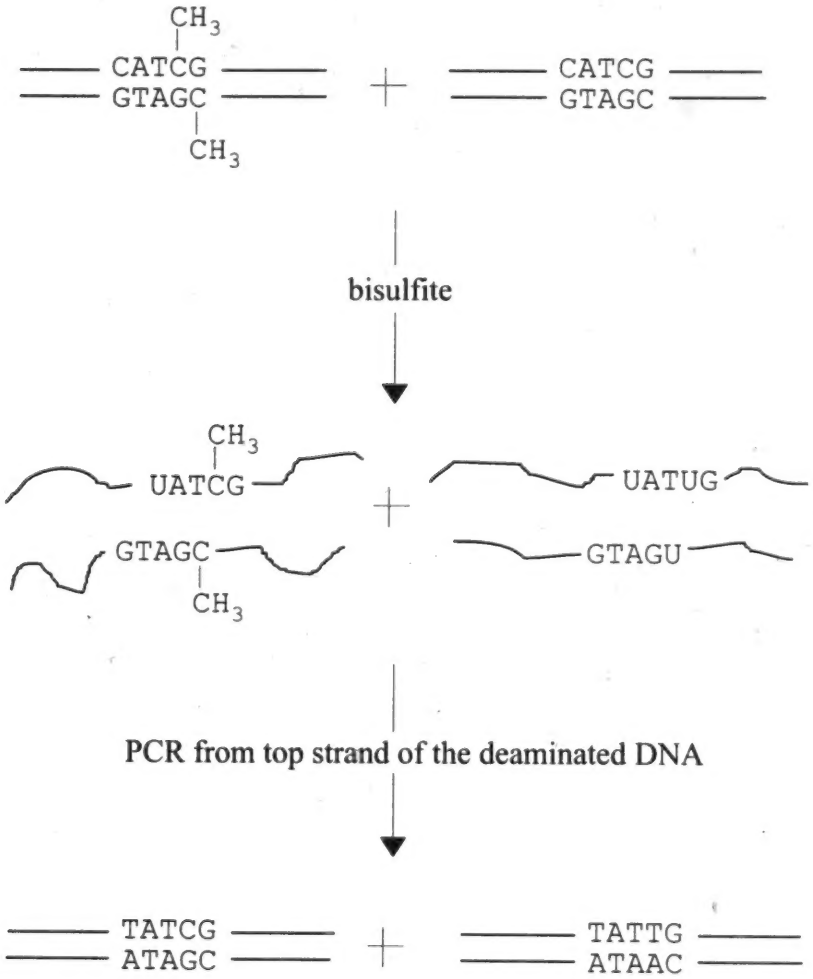
PCR primers will anneal to a three-base code, A, G, and T or U. The DNA polymerase used in the PCR will replicate U as T and 5-Me-C as C. In the final PCR product, positions that correspond to unmethylated cytosine in the genomic DNA will now be T and positions that correspond to a 5-Me-C will now be C. The methylation state of each cytosine in this DNA sequence is determined by sequencing the PCR product (known as a CD-PCR product) with standard DNA sequencing techniques.

The CD-PCR technology was initially developed by Frommer *et al.*[125] and has undergone several improvements since its introduction[126-132]. The CD-PCR technology can be extraordinarily sensitive: genomic DNA from as few as 100 cells has been analyzed[127, 129]. Because CD-PCR technology does not require fragmenting the genomic DNA to obtain a sequencing ladder, the theoretical limit of its sensitivity is one molecule. Further, because CD-PCR chemistry is able to select individual strands of the genomic DNA for analysis it is able to detect hemi-methylated CpG sites.

The bisulfite-based deamination step in CD-PCR chemistry requires single-stranded DNA. Double-stranded DNA will not fully deaminate, leading to sections of DNA where unmethylated cytosines are not converted to uracil. This effect destroys the methylation analysis data in this area. Some DNA sequences appear to be more resistant to deamination with sodium bisulfite than other sequences[133], likely because these sequences are difficult to fully denature.

If the original genomic DNA sample has cytosines which are partially methylated, its CD-PCR product will contain more than one sequence (a degenerate CD-PCR product sequence). Figure 1.14 shows how a CD-PCR product with a degenerate sequence can arise from genomic DNA differing only in the positions of methylated cytosines. Obtaining the sequence of degenerate CD-PCR products can be done in two ways. The CD-PCR products may be cloned and several clones sequenced; this is equivalent to determining the sequence of a single CD-PCR product molecule, which was derived from a single chromosome molecule in the genomic DNA. Cloning CD-PCR products makes possible the determination of methylation status of individual chromosomes. The other option is to sequence the CD-PCR products directly. Direct sequencing of CD-PCR products may provide the average methylation status of a DNA sequence in a single

Figure 1.14 Creation of a degenerate CD-PCR from genomic DNA that differs only in methylation state



analysis. Chapter 4 describes the development of a system that measures the average methylation status of partially methylated DNA samples in a single analysis by direct cycle-sequencing of CD-PCR products. Cloning versus direct sequencing of CD-PCR products is also discussed in more detail in Chapter 4.

CD-PCR chemistry has the potential to produce quantitative information about partially methylated cytosines. This depends entirely upon the ability of the PCR to evenly amplify the DNA from the methylated (the left sequence in Figure 1.14) and deaminated genomic DNA with respect to the unmethylated and deaminated DNA (the right sequence in Figure 1.14). This can by no means be taken for granted: slight differences in the amplification efficiency of the sequences from deaminated DNA with different methylation states will invalidate any quantitative results. The significance of this problem and a solution to it are discussed in Chapter 5.

1.4 Bibliography

1. Sanger, F.; Nicklen, S. and Coulson, A. R. *Proceedings of the National Academy of Sciences, USA* **74**, 5463-5467 (1977).
2. Smith, L. M.; Sanders, J. Z.; Kaiser, R. J.; Huges, P.; Dodd, C.; Connel, C. R.; Heiner, C.; Kent, S. B. H. and Hood, L. E. *Nature* **321**, 674-679 (1986).
3. Prober, J. M.; Trainor, G. L.; Dam, R. J.; Hobbs, F. W.; Robertson, C. W.; Zagursky, R. J.; Cocuzza, A. J.; Jensen, M. A. and Baumeister, K. *Science* **238**, 336-341 (1987).
4. Tiselius, A. *Nova Acta Regiae Societatis Scientiarum Upsaliensis* (1930).
5. Tiselius, A. *Transactions of the Faraday Society* **33**, 524-531 (1937).
6. Raymond, S. and Weintraub, L. *Science* **130**, 711 (1959).
7. Hjerten, S. *Biochemistry and Biophysic Acta* **53**, 514-517 (1961).
8. Barron, A. E. and Blanch, H. W. *Separation and Purification Methods* **24**, 1-118 (1995).
9. Tietz, D. in *Advances in Electrophoresis* (eds. Chrambach, A., Dunn, M.J. & Radola, B.J.) (VCH, Weinheim, 1988).
10. Viovy, J. L. and Heller, C. in *Capillary Electrophoresis in Analytical Biotechnology* (ed. Righetti, P.G.) (CRC Press, Boca Raton, 1996).
11. Righetti, P. G. and Gelfi, C. in *Capillary Electrophoresis in Analytical Biotechnology* (ed. Righetti, P.G.) (CRC Press, Boca Raton, 1996).
12. Slater, G. W. in *Analysis of Nucleic Acids by Capillary Electrophoresis* (ed. Heller, C.) (Friedr. Vieweg &Sohn Verlagsgesellschaft, Braunschweig, 1997).
13. Ogston, A. G. *Trans. Faraday Society* **54**, 1754-1757 (1958).
14. Rodbard, D. and Chrambach, A. *Proceedings of the National Academy of Sciences, USA* **65**, 970-977 (1970).
15. Holmes, D. L. and Stellwagen, N. C. *Electrophoresis* **11**, 5-15 (1990).
16. Heller, C.; Duke, T. and Viovy, J. L. *Biopolymers* **34**, 249-259 (1994).
17. Jan, J. Y.; Best, N.; Zhang, J. Z.; Ren, H. J.; Jiang, R.; Hou, J. and Dovichi, N. J. *Electrophoresis* **17**, 1037-1045 (1996).
18. Fangmann, W. L. *Nucleic Acids Research* **5**, 653-655 (1978).

19. Slater, G. W. and Noolandi, J. *Biopolymers* **25**, 431-454 (1986).

20. Slater, G. W.; .Turmel, C.; Lalande, M. and Noolandi, J. *Biopolymers* **28**, 1793-1799 (1989).

21. Duke, T.; Viovy, J. L. and Semenov, A. N. *Biopolymers* **34**, 239-248 (1994).

22. Slater, G. W. and Drouin, G. *Electrophoresis* **13**, 574-582 (1992).

23. Slater, G. W. *Electrophoresis* **14**, 1-7 (1993).

24. Zhang, J. Z.; Fang, Y.; Ren, H. J.; R.Jiang; Roos, P. and Dovichi, N. J. *Analytical Chemistry* **67**, 4589-4593 (1995).

25. Dovichi, N. J. and Zhang, J. Z. United States Patent, No. 5584982 (1996).

26. Hoagland, D. A. and Arvantidou, E. S. *Polymer Preprints* **34**, 1059-1060 (1993).

27. Cohen, A. S.; Najarian, D. R.; Paulus, A.; Guttman, A.; Smith, J. A. and Karger, B. L. *Proceedings of the National Academy of Sciences, USA* **85**, 9660-9703 (1988).

28. Swerdlow, H.; Harke, H. R.; Wu, S. and Dovichi, N. J. *Journal of Chromatography* **516**, 61-67 (1990).

29. Swerdlow, H. and Gesteland, R. *Nucleic Acids Research* **18**, 1415-1420 (1990).

30. Drossman, H.; Luckey, J. A.; Kostichka, A. J.; D'Cunha, J. and Smith, L. M. *Analytical Chemistry* **62**, 900-903 (1990).

31. Cohen, A. S.; Najarian, D. R. and Karger, B. L. *Journal of Chromatography* **516**, 49-60 (1990).

32. Swerdlow, H.; Dew-Jager, K. E.; Grey, R.; Dovichi, N. J. and Gesteland, R. *Electrophoresis* **13**, 475-483 (1992).

33. Bode, H. J. *Analytical Biochemistry* **83**, 204-210 (1977).

34. Bode, H. J. *Analytical Biochemistry* **83**, 364-371 (1977).

35. Tietz, D.; Aldroubi, A.; Pulyaeva, H.; Guszczynski, T.; Garner, M. M. and Chrambach, A. *Electrophoresis* **13**, 614-616 (1992).

36. Pulyaeva, H.; Wheeler, D.; Garner, M. M. and Chrambach, A. *Electrophoresis* **13**, 608-614 (1992).

37. Tietz, D.; Gottlieb, M. H.; Fawcett, J. S. and Chrambach, A. *Electrophoresis* **7**, 217-220 (1986).

38. Best, N.; Arriaga, E.; Chen, D. Y. and Dovichi, N. J. *Analytical Chemistry* **66**, 4063-4067 (1994).

39. Ruiz-Martinez, M. C.; Berka, J.; Belenkii, A.; Foret, F.; Miller, A. W. and Karger, B. L. *Analytical Chemistry* **65**, 2851-2858 (1993).

40. Fung, E. N. and Yeung, E. S. *Analytical Chemistry* **67**, 1913-1919 (1995).

41. Menchen, S.; Johnson, B.; Winnik, M. A. and Xu, B. *Electrophoresis* **17**, 1451-1459 (1996).

42. Sudor, J.; Foret, F. and Bocek, P. *Electrophoresis* **12**, 1056-1058 (1991).

43. Grossman, P. D. and Soane, D. S. *Biopolymers* **31**, 1221-1228 (1991).

44. Viovy, J. L. and Duke, T. *Electrophoresis* **14**, 322-329 (1993).

45. Broseta, D.; Leibler, L.; Lapp, A. and Strazielle, C. *Europhys. Letters* **2**, 733-737 (1986).

46. Carriho, E.; Ruiz-Martinez, M. C.; Berka, J.; Smirnov, I.; Goetzinger, W.; Miller, A. W.; Brady, D. and Karger, B. L. *Analytical Chemistry* **68**, 3305-3313 (1996).

47. Figeys, D.; Renborg, A. and Dovichi, N. J. *Electrophoresis* **15**, 1512-1517 (1994).

48. Figeys, D. and Dovichi, N. J. *Journal of Chromatograpy* **717**, 105-111 (1995).

49. Figeys, D.; Ahmadzedeh, H.; Arriaga, E. and Dovichi, N. J. *Journal of Chromatography* **744**, 325-331 (1996).

50. Figeys, D. and Dovichi, N. J. *Journal of Chromatography* **744**, 333-339 (1996).

51. Gelfi, C.; Desi, P. d.; Alloni, A. and Righetti, P. G. *Journal of Chromatography* **608**, 333-341 (1992).

52. Chiari, M.; Nesi, M. and Righetti, P. G. *Electrophoresis* **15**, 616-622 (1994).

53. Chiari, M.; Micheletti, C.; Nesi, M.; Fazio, M. and Righetti, P. G. *Electrophoresis* **15**, 177-186 (1995).

54. Simo-Alfonso, E.; Gelfi, C.; Sabastiano, R.; Citterio, A. and Righetti, P. G. *Electrophoresis* **17**, 723-731 (1996).

55. Simo-Alfonso, E.; Gelfi, C.; Sabastiano, R.; Citterio, A. and Righetti, P. G. *Electrophoresis* **17**, 732-737 (1996).

56. Gelfi, C.; Simo-Alfonso, E.; Sabastiano, R.; Citterio, A. and Righetti, P. G. *Electrophoresis* **17**, 738-743 (1996).

57. Hjerten, S. J. *Journal of Chromatography* **347**, 191-198 (1985).

58. Belder, D. and G.Schomburg. *Journal of High Resolution Chromatography* **15**, 686-693 (1992).

59. Kleemiss, M. H.; Giles, M. and Schomburg, G. *Electrophoresis* **14**, 515-522 (1993).

60. Madabhushi, R. S.; Menchen, S. M.; Efcavitch, J. W. and Grossman, P. D. United States Patent, No. 5552028 (1996).

61. Giles, M.; Kleemiss, M. H. and Schomburg, G. *Analytical Chemistry* **66**, 2038-2046 (1994).

62. Busch, M. H. A.; Kraak, J. C. and Poppe, H. *Journal of Chromatography* **695**, 287-296 (1995).

63. Schmalzing, D.; Pigee, C. A.; Foret, F.; Carrilho, E. and Karger, B. L. *Journal of Chromatography* **652**, 149-159 (1993).

64. Cobb, K. A.; Dolnik, V. and Novotny, M. *Analytical Chemistry* **62**, 2478-2483 (1990).

65. Konrad, K. D. and Pentoney, S. L. *Electrophoresis* **14**, 502-508 (1993).

66. Fang, Y.; Zhang, J. Z.; Hou, J. Y.; Lu, H. and Dovichi, N. J. *Electrophoresis* **17**, 1436-1442 (1996).

67. Hotchkiss, R. D. *Journal of Biological Chemistry* **168**, 315-332 (1948).

68. Watson, J. D. and Crick, F. H. C. *Nature* **171**, 964-967 (1953).

69. Watson, J. D. and Crick, F. H. C. *Nature* **171**, 737-738 (1953).

70. Li, E.; Bestor, T. H. and Jaenisch, R. *Cell* **69**, 915-926 (1992).

71. Schwartz, M. N.; Trautner, T. A. and Kornberg, A. *Journal of Biological Chemistry* **237**, 1961-1967 (1962).

72. Bestor, T. H.; Hellwell, S. B. and Ingram, V. *Molecular Cell Biology* **4**, 1800-1806 (1984).

73. Clark, S. J.; Harrison, J. and Frommer, M. *Nature Genetics* **10**, 20-27 (1995).

74. Gruenbaum, Y. *Proceedings of the National Academy of Sciences, USA* **80**, 4919-4921 (1983).

75. Smith, S. S. *Progress in Nucleic Acids Research and Molecular Biology* **49**, 65-

111 (1994).

76. Yen, R. C.; Vertino, P. M.; Nelkin, B. D.; Yu, J. J.; Wafik, E.; Cumaraswarmy, A.; Lennon, G. G.; Trask, B. J.; Celano, P. and Baylin, S. B. *Nucleic Acids Research* **20**, 2287-2291 (1992).

77. Yoder, J. A.; Yen, R. C.; Vertino, P. M.; Bestor, T. H. and Baylin, S. B. *The Journal of Biological Chemistry* **271**, 31092-31097 (1996).

78. Bestor, T. H. and Verdine, G. L. *Current Opinions in Cell Biology* **6**, 380-389 (1994).

79. Chuang, L. S. H.; Ng, H.; Chia, J. and Li, B. F. L. *The Journal of Molecular Biology* **257**, 935-948 (1996).

80. Layoun, A. and Smith, S. S. *Nucleic Acids Research* **23**, 1584-1589 (1995).

81. Murchie, A. I. H. and Lilley, D. M. J. *Nucleic Acids Research* **15**, 9641-9654 (1989).

82. Hagerman, P. J. *Biochemistry* **29**, 1980-1983 (1990).

83. Diekmann, S. *EMBO Journal* **6**, 4213-4217 (1987).

84. Lee, J. S.; Woodsworth, M. L.; Latimer, L. J. P. and Morgan, A. R. *Nucleic Acids Research* **12**, 6603-6614 (1984).

85. Rich, A.; Nordheim, A. and Wang, A. H. J. *Annual Reviews in Biochemistry* **53**, 791-846 (1984).

86. Behe, M. and Felsenfeld, G. *Proceedings of the National Academy of Sciences, USA* **78**, 1619-1623 (1981).

87. Antequera, F. and Bird, A. in *DNA Methylation: Molecular Biology and Biological Significance* (eds. Jost, J.P. & Saluz, H.P.) 167-185 (Birkhauser, Bassel, 1993).

88. Yeivin, A. and Razin, A. in *DNA Methylation: Molecular Biology and Biological Significance* (eds. Jost, J.P. & Saluz, H.P.) 523-568 (Birkhauser, Bassel, 1993).

89. Prendergast, G. C.; Lawe, D. and Ziff, E. B. *Cella* **65**, 395-407 (1991).

90. Inamdar, N. M.; Ehrlich, K. C. and Ehrlich, M. *Plant Molecular Biology* **17**, 111-123 (1991).

91. Zhang, X. Y.; Asiedu, C. K.; Supakar, P. C.; Khan, R.; Ehrlich, K. C. and Ehrlich,

M. *Nucleic Acids Research* **18**, 6253-6260 (1990).

92. Meehan, R. R.; Lewis, J. D.; McKay, S.; Kleiner, E. L. and Bird, A. P. *Cell* **58**, 499-507 (1989).

93. Lewis, J. and Bird, A. *FEBS letters* **285**, 155-159 (1991).

94. Christophe, D. and Pichon, B. *Molecular and Cellular Endocrinology* **100**, 155-158 (1994).

95. Monk, M. *Developmental Genetics* **17**, 188-197 (1995).

96. Grant, S. G. and Chapman, V. M. *Annual Reviews in Genetics* **22**, 199-233 (1988).

97. Duncan, B. K. and Miller, J. H. *Nature* **287**, 560-561 (1980).

98. Harris, C. C. and Hollstein, M. *New England Journal of Medicine* **329**, 1318-1327 (1993).

99. Laird, P. W. and Jaenisch, R. *Annual Review of Genetics* **30**, 441-464 (1996).

100. Rao, P. M.; Antony, A.; Rajalakshmi, S. and Sarma, D. S. *Carcinogenesis* **10**, 933-937 (1989).

101. Ray, J. S.; Harbison, M. L.; McClain, R. M. and Goodman, J. I. *Molecular Carcinogenesis* **9**, 155-166 (1994).

102. Vorce, R. L. and Goodman, J. I. *Toxicology and Applied Pharmacology* **100**, 398-410 (1989).

103. Sakai, T.; Toguchida, J.; Ohtani, N.; Yandell, D. W.; Rapaport, J. M. and Dryja, T. P. *American Journal of Human Genetics* **48**, 880-888 (1991).

104. Greger, V.; Passarge, E.; Hopping, W.; Messmer, E. and Horsthemke, B. *Human Genetics* **83**, 155-158 (1898).

105. Herman, J. G.; Latif, F.; Weng, Y.; Lerman, M. I.; Zbar, B.; Liu, S.; Samid, D.; Duan, D. S.; Gnarra, J. R. and Linehan, W. M. *Proceedings of the National Academy of Sciences, USA* **91**, 9700-9704 (1994).

106. Merlo, A.; Herman, J. G.; Mao, L.; Lee, D. J.; Gabrielson, E.; Burger, P. C.; Balin, S. B. and Sidransky, D. *Nature Medicine* **1**, 686-692 (1995).

107. Gonzalez-Zulueta, M.; Bender, C. M.; Yang, A. S.; Nguyen, T.; Beart, R. W.; Tornout, J. M. B. and Jones, P. A. *Cancer Research* **55**, 4531-4535 (1995).

108. Herman, J. G.; Merlo, A.; Mao, L.; Lapidus, R. G.; Issa, J.-P. J.; Davidson, N. E.; Sidranski, D. and Baylin, S. B. *Cancer Research* **55**, 4525-4530 (1995).

109. Rideout, W. M.; Eversole-Cire, P.; Spruck, C. H.; Hustad, C. M.; Coetzee, G. A.; Gonzales, F. and Jones, P. A. *Molecular Cell Biology* **14**, 6143-6152 (1994).

110. Yoshiura, K.; Kanai, Y.; Ochiai, A.; Shimoyama, Y.; Sugimura, T. and Hirohashi, S. *Proceedings of the National Academy of Sciences, USA* **92**, 7416-7419 (1995).

111. Tollefsbol, T. O. and Hutchinson, C. A. *Journal of Biological Chemistry* **270**, 18543-18550 (1995).

112. Tollefsbol, T. O. and Andrews, L. G. *Medical Hypothesis* **41**, 83-92 (1993).

113. Ledoux, S.; Nalbantoglu, J. and Cashman, N. R. *Molecular Brain Research* **24**, 140-144 (1994).

114. Rogaev, E. I.; Lukiw, W. J.; Lavrushina, O.; Rogaeva, E. A. and George-Hyslop, P. H. S. *Genomics* **22**, 340-347 (1994).

115. Bird, A. P. and Southern, E. M. *Journal of Molecular Biology.* **118**, 27-47 (1978).

116. Singer-Sam, J.; Yang, T. P.; Mori, N.; Tanguay, R. L.; Le-Bon, J. M.; Flores, J. C. and Riggs, A. D. in *Nucleic Acid Methylation* (eds. Clawson, G.A., Willis, D.B., Weissbach, A. & Jones, P.A.) 285-289 (Alan R. Liss, Inc., New York, 1990).

117. Maxam, A. M. and Gilbert, W. *Methods in Enzymology* **65**, 499-560 (1984).

118. Church, G. M. and Gilbert, W. *Proceedings of the National Academy of Sciences, USA* **81**, 1991-1995 (1984).

119. Saluz, H. P. and Jost, J. P. *Gene* **42**, 151-157 (1986).

120. Pfeifer, G. P.; Steigerwald, S. D.; Mueller, P. R.; Wold, B. and Riggs, A. D. *Science* **246**, 810-813 (1989).

121. Hornstra, I. K. and Yang, T. P. *Molecular and Cellular Biology* **14**, 1419-1430 (1994).

122. Hayatsu, H.; Yusuke, W.; Kazushige, K. and Shigeru, I. *Biochemistry* **9**, 2858-2865 (1970).

123. Wang, R. Y.; Gehrke, C. W. and Ehrlich, M. *Nucleic Acids Research* **8**, 4777-4790 (1980).

124. Shapiro, R.; DiFate, V. and Welcher, M. *Journal of the American Chemical*

Society **96**, 906-912 (1974).

125. Frommer, M.; McDonald, L. E.; Millar, D. S.; Collis, C. M.; Watt, F.; Grigg, G. W.; Molloy, P. L. and Paul, C. L. *Proceedings of the National Academy of Sciences, USA* **89**, 1827-1831 (1992).

126. Feil, R.; Charlton, J.; Bird, A. P.; Walter, J. and Reik, W. *Nucleic Acids Research* **22**, 695-696 (1994).

127. Clark, S. J.; Harrison, J.; Paul, C. and Frommer, M. *Nucleic Acids Research* **22**, 2990-2997 (1994).

128. Myohanen, S.; Wahlfors, J. and Janne, J. *DNA Sequence* **5**, 1-8 (1994).

129. Olek, A.; Oswald, J. and Walter, J. *Nucleic Acids Research* **24**, 5064-5066 (1996).

130. Paul, C. L. and Clark, S. J. *BioTechniques* **21**, 126-133 (1996).

131. Raizis, A. M.; Schmitt, F. and Jost, J.-P. *Analytical Biochemistry* **226**, 161-166 (1995).

132. Reeben, M. and Prydz, H. *BioTechniques* **16**, 416-417 (1994).

133. Rother, K. I.; Silke, J.; Georgiev, O.; Schaffner, W. and Matsuo, K. *Analytical Biochemistry* **231**, 263-265 (1995).

CHAPTER 2: DNA SEQUENCING WITH POLY N,N-DIMETHYLACRYAMIDE: STABILITY OF THE CAPILLARY COATING

2.1 Introduction

As discussed in Chapter 1, DNA sequencing by capillary electrophoresis has the potential to increase total sequencing throughput. While electrophoretic separation times for DNA sequencing applications by capillary electrophoresis are similar to competing slab gel systems, the primary advantages of capillary electrophoresis are the possibility of multi-capillary instrumentation with upwards of 1000 capillaries and the ease of with which the instrumentation can be automated.

The original CE DNA sequencing separations used a cross-linked polyacrylamide gel as a sieving matrix[1-3]. Such capillaries are not useable indefinitely: they must be replaced after a few sequencing separations. Unfortunately, it is not feasible to manually replace the capillaries in an automated CE instrument. It would be preferable to replace the sieving matrix inside the capillary instead of the capillary itself. Non-crosslinked polyacrylamide has been used as a replaceable sieving matrix[4-6]. Other polymer solutions such as polyethylene oxide[7] and modified polyethylene glycol[8] have also been used as a sieving matrix for DNA sequencing. However, these studies either do not achieve sufficient read length (500-600 bases) or do not report how many times the capillaries are actually used.

The failure of DNA sequencing separations performed when replacing the sieving matrix could be due to a destructive change in the sieving matrix or to destruction of the capillary wall coating (see Section 1.2.6.3). Capillaries are commonly coated by the method of Hjerten[9]. Righetti and coworkers have proposed that alkaline hydrolysis of both the polyacrylamide sieving matrix and the polyacrylamide layer of the Hjerten coating may be significant[10]. They have developed polymers made from N-substituted acrylamide derivatives that are both highly hydrophilic and stable to alkaline hydrolysis. Novotny and coworkers have proposed that alkaline hydrolysis of the silane linkages in the Hjerten type capillary coating can destroy the coating over time[11]. They have developed a coating procedure that uses a Grignard coupling procedure to avoid the use of labile silicon-oxygen bonds in attaching the coating to the capillary wall. However, one study has shown that 30 hours at pH 10 is required to cause significant hydrolysis of the silicon-oxygen bonds in the Hjerten coating[12].

I wished to confirm that failure of the capillary coating was responsible for poor sequencing separations obtained in successive runs with replaced sieving matrix. I also wished to determine if alkaline hydrolysis was indeed responsible for coating failure. Finally, I wished to determine if poly-N,N-dimethylacrylamide (DMA) would function as a sieving matrix for DNA sequencing.

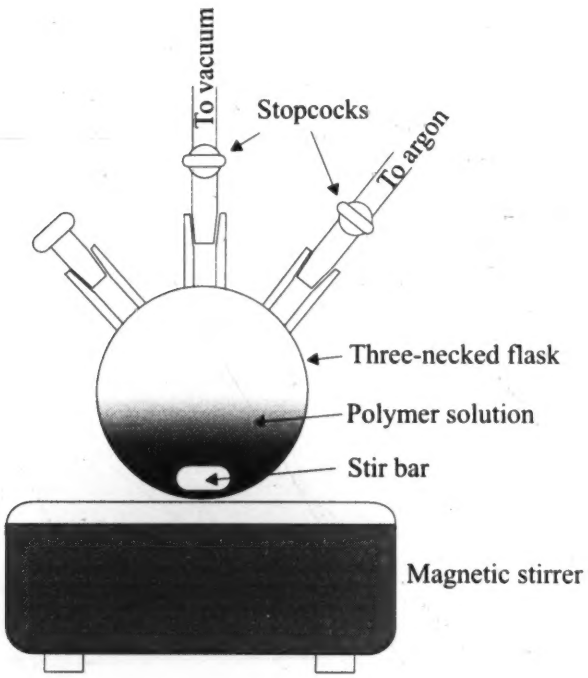
To determine if DMA is suitable for DNA sequencing by CE, coated capillaries were filled with poly-DMA and used for DNA sequencing. To determine if hydrolysis of the capillary coating was responsible for failures of the sequencing runs, a capillary coated by the Hjerten procedure was used it for successive sequencing runs using poly-DMA as the replaceable sieving matrix. As a control, 5 capillaries coated by the Hjerten method were filled with sieving matrix at the same time and one capillary was used for each subsequent run. Finally, capillaries coated by the Novotny procedure were tested to determine if this coating was more robust and suitable for DNA sequencing applications.

2.2 Experimental

2.2.1 Preparation of poly-N,N-dimethylacrylamide

A 6% poly-DMA solution for use as a sieving matrix was made as follows. 2 mL of 5X TBE buffer (445 mM tris, 445 mM borate, 50 mM EDTA), 4.2 g urea (ICN) and 620 µL of DMA (Sigma) were dissolved in 10 mL of water. The solution was passed through a 0.22 µm filter (Millipore) into a 50 mL three-necked flask. A diagram of the apparatus is shown in Figure 2.1. The stopcock to the argon was closed and the solution degassed by vacuum for 5 minutes while stirring. The amount of water that evaporated during the degassing procedure was measured by weighing the flask before and after degassing. Approximately 0.25 mL of water evaporated during degassing. After degassing, the vacuum stopcock was closed, the stirbar stopped, and the argon stopcock was opened to fill the flask with argon gas. Argon was kept flowing while the stopper was removed and 20 µL of 10% ammonium persulfate solution was added and the solution stirred for a few seconds. To initiate polymerization, 10 µL of TEMED was added, the solution stirred, and the stopper replaced. The stopcock to the argon was closed and the solution was left to polymerize overnight at room temperature. After polymerization the solution was poured into a 10 mL syringe and stored at 4°C. The viscosity of the polymer

Figure 2.1 Apparatus used for the polymerization of poly-DMA



solution was measured with a falling ball viscometer and determined to be 3100 cp at room temperature. For comparison, glycerol has a viscosity of 1500 cp. One batch of polymer was used for all experiments in this chapter.

2.2.2 Preparation of T-terminated sequencing samples

T-terminated sequencing samples were prepared by cycle sequencing. 1.5 µL of M13mp18 single-stranded DNA (Amersham), 2 µL of M13-21 Rox-labeled primer (Perkin Elmer), 4 µL of Thermosequenase™ T reagent (Amersham) and 8.5 µL of water were mixed in a 200 µL PCR tube. Thermal cycling was done with a MJ Research PTC-100 with heated lid. The cycling conditions were as follows: 30 cycles of 95 °C for 30 seconds and 55 °C for 30 seconds. Two identical sequencing samples were prepared simultaneously. After cycling, both reactions were combined in a 600 µL centrifuge tube, precipitated with 400 µL of 95% ethanol and resuspended in 30 µL of deionized formamide. The sequencing sample was divided into 5 µL aliquots for injection onto the capillary electrophoresis instrument. Samples were stored at -20 °C until used.

2.2.3 Preparation of silane-coated capillaries (Hjerten coating)

A silanizing solution containing 10 µL of γ-methacryloxypropyltrimethoxysilane (Sigma) in 500 uL of 95% ethanol was drawn through capillaries (150 µm O.D. x 50 µm I.D. x 43 cm long) by vacuum for 30 minutes followed by air for 10 minutes. The silanized capillaries were then filled with a polymerizing 4% DMA, 1X TBE, and 7 M urea solution that was initiated with 10 µL of 10% ammonium persulfate (BioRad) and 2 µL of TEMED (BioRad). The coated capillaries were stored with the polymer in them until used.

2.2.4 Preparation of Grignard coated capillaries (Novotny coating)

The tetrahydrofuran and vinyl magnesium bromide used for this procedure were purchased from Sigma in SureSeal™ bottles and aliquots were removed with a syringe to ensure that the reagents remained free of moisture. A four-meter length of capillary was flushed with 1 M NaOH for 3 hours, followed by water for 1 hour and methanol for 1 hour. Flushing was done by sealing one end of the capillary in a vial that contained the reagent to be flushed through the capillary. The vial was connected to a high pressure nitrogen tank, which allowed me to bring the nitrogen atmosphere inside the vial to any

desired pressure. Reagents were forced through the capillary by applying a 20 psi nitrogen atmosphere to the vial. After the methanol flush the capillary was flushed with dry nitrogen at 5 psi overnight in a 120 °C oven. The next day the oven temperature was reduced to 65 °C. Thionyl chloride (Sigma), 0.5 mL, was flushed through the capillary for 30 minutes by applying 20 psi of nitrogen. Both ends of the capillary were capped with a GC septum and the capillary incubated for 8 hours at 65 °C. After incubation, one freshly cut end of the capillary was placed in a vial containing 1.0 mL of THF and 0.2 mL of 2 M vinylmagnesium bromide. The other end of the capillary was placed in methanol. The THF/vinylmagnesium bromide solution was flushed through the capillary for 30 minutes by applying 20 psi nitrogen, the capillary tips capped with GC septa, and the capillary incubated for 8 hours at 65 °C. After incubation, the capillary was rinsed with anhydrous THF for 30 minutes by applying 20 psi nitrogen, followed by a 30 minute rinse with water.

To complete the coating, the functionalized capillary was filled with a polymerizing 4% N,N-dimethylacrylamide solution in the same manner as described for the silane functionalized capillary (Section 2.2.3). Following polymerization overnight, the capillaries were ready to be filled with sieving matrix and used for DNA sequencing.

2.2.5 Instrumentation

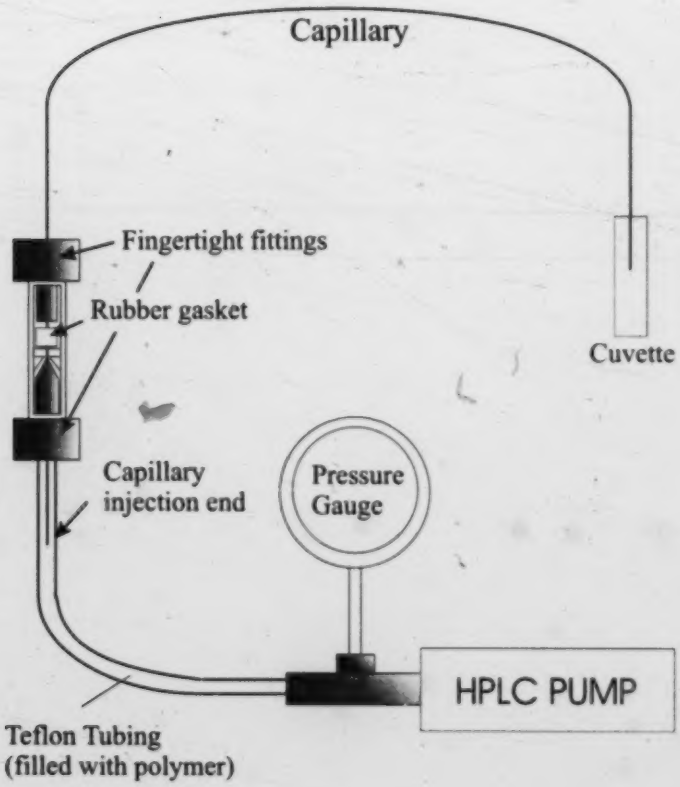
A locally constructed single capillary instrument with sheath flow cuvette was used for all the sequencing separations[13]. This instrument was configured to use the 543 nm line of the helium-neon laser for excitation and a 640DF40 bandpass filter (Omega Optical) and R1477 PMT (Hamamatsu) for detection.

2.2.6 Sequencing separations

2.2.6.1 Experiment 1: Multiple runs on a single capillary

A one day old capillary coated by the Hjertein method (section 2.2.3) was placed in the single-capillary electrophoresis instrument. The capillary was filled with the poly-DMA matrix with a high pressure Isco model 316 HPLC pump. This apparatus is detailed in Figure 2.2. A short section of Teflon tubing was filled with poly-DMA. One end of the tubing was attached to the syringe-type HPLC pump with fingertight fittings (Upchurch Scientific). The injection end of the capillary was sealed in the other end of the tubing

Figure 2.2 System used to replace the poly-DMA polymer solution



with a rubber gasket cut from a rubber stopper, fingertight fittings and a HPLC union. The pump was manually cranked to a pressure of 1000 psi. It was not necessary to activate the pump motor as the total inner volume of a capillary is approximately 3 μL; no significant pressure drop occured as the capillary was being filled. With this system an empty capillary could be filled in less than 20 minutes. To ensure that the sieving matrix was completely replaced, the polymer was pushed through the capillary for 1 hour at each refill. After each refill a 190 V/cm reverse polarity electric field (cathode at the detection end) was applied to the capillary for two minutes. Before injection the capillaries were run with normal polarity for 10 minutes at 190 V/cm to allow the current to stabilize. A small aliquot of poly-DMA sieving matrix as described in Section 2.2.1, containing TBE buffer and urea, was used as running buffer at the injection end. TBE buffer was used in the sheath flow. Just before injection, capillaries were run with reverse polarity at 100V/cm for 30 seconds in a TBE-7 M urea running buffer. Samples were injected for 40 seconds at a 100 V/cm field, the poly-DMA running buffer was replaced and the separations were performed at a field of 190 V/cm. Fluorescence intensity was sampled at 2 Hz. Four runs were performed on this capillary, one run each consecutive day.

2.2.6.2 Experiment 2: Single runs on aged capillaries (Hjerten coating)

Instead of reusing the same capillary for 5 runs a new capillary was used for each run. Five capillaries were coated, as described in section 2.2.3, one day previous to the start of this experiment. The capillary was again filled with sieving matrix for one hour; injection and running conditions were identical to those described in Section 2.2.6.1. However, instead of reusing a single capillary, the capillaries were filled with polymer once, used for a single separation, and replaced the next day. Five runs were performed, one run each consecutive day.

2.2.6.3 Experiment 3: Multiple runs on a Grignard-coated capillary

A single capillary was coated as described in section 2.2.4. This capillary was used for multiple sequencing runs with the poly-DMA sieving matrix. Refilling and running conditions were identical to those in Section 2.2.6.1 except that 10 runs were performed instead of 4.

2.2.3 Data analysis

Electropherograms were imported into PeakFit v 4.0 (Jandel Scientific) and Gaussian curves fit to each peak. Curves were only fit to peaks that were recognizable as belonging to a certain fragment length(s). Curves were not fit to peaks later in the runs with very poor resolution because it was difficult to distinguish individual peaks or groups of peaks. Normalized resolution was calculated according the equation:

$$Normalized\ Resolution = 2\times\frac{t_2 - t_1}{W_1 + W_2}\times\frac{1}{M_2 - M_1} \qquad (2.1)$$

where t_1 and t_2 are the migration times of the two peaks and W_1 and W_2 are the full widths at the base of the peaks. M_1 and M_2 are the size of the DNA fragments, in bases, corresponding to the two peaks. Calculation of the number of theoretical plates was according to the equation:

$$N = 5.5\left(\frac{t}{W_{1/2}}\right)^2 \qquad (2.2)$$

where t is the migration time for a peak and $W_{1/2}$ is the width at half maximum of the same peak.

2.3 Results and Discussion

2.3.1 Suitability of poly-DMA sieving matrix

Figure 2.3 shows the electropherogram from the first run of Experiment 1. Resolution is good to over 500 bases. This is the first report of poly-DMA being used for sequencing separations. The relatively low viscosity of the poly-DMA solution at 3100 cp allows it to be easily pumped through the capillaries with a high pressure pump. Like the substituted acrylamides developed by Righetti[14-17] DMA is highly resistant to alkaline hydrolysis[10]. Unlike Righetti's monomers, which require considerable effort to synthesize, DMA can be purchased from commercial suppliers.

Figure 2.4 is the electropherogram of the second run on the same capillary. The resolution in this run is good only to approximately 320 bases. Qualitatively, there is very little difference in the electropherogram in Figure 2.4 from the electropherogram in

Figure 2.3 Electropherogram of a T-termination sequencing reaction. First run on this capillary. Peaks are for bases 26 to 538

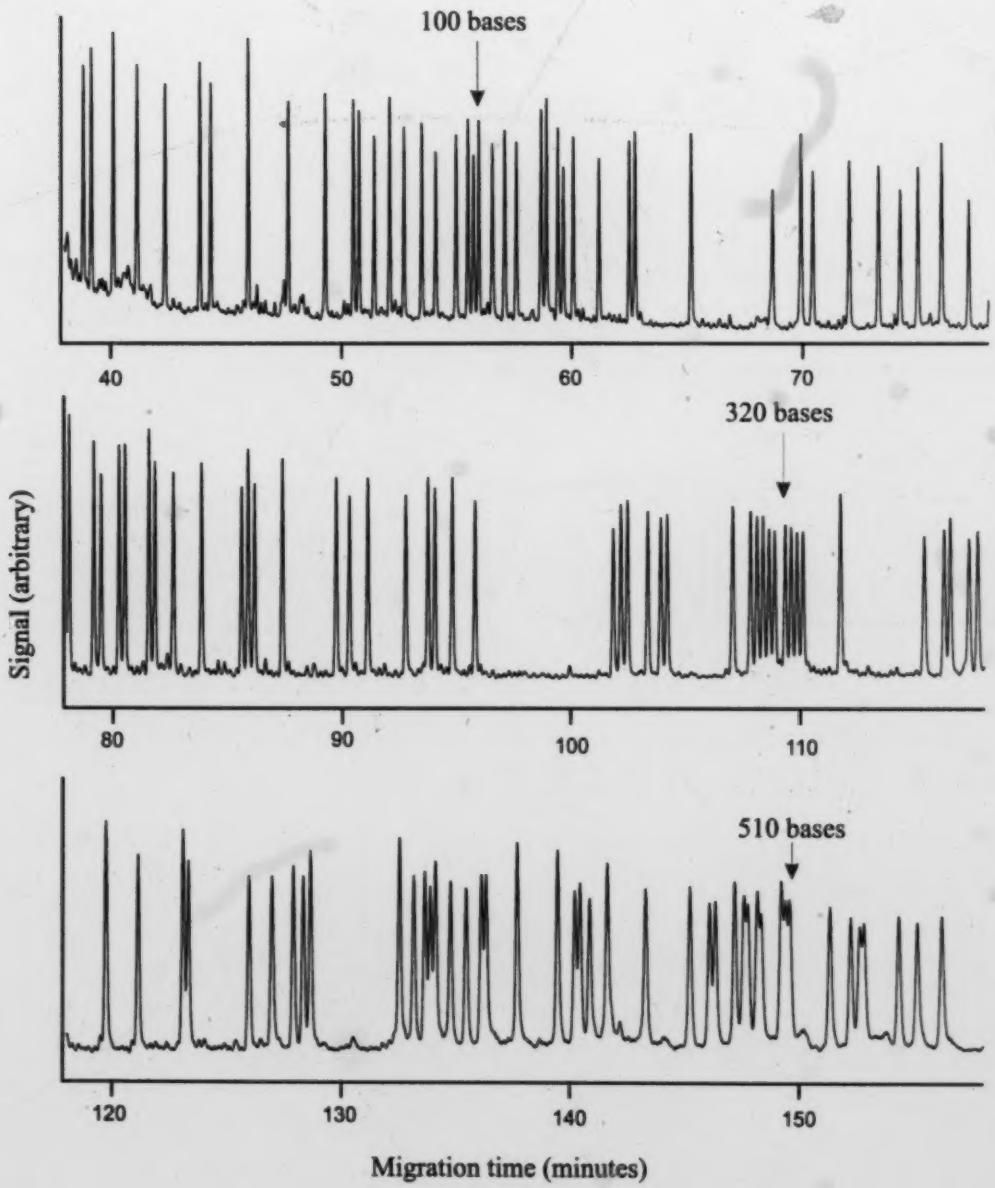


Figure 2.4 Electropherogram of T-termination sequencing reaction. This is the second run on this capillary after refilling with polymer. Peaks are for bases 26 - 538

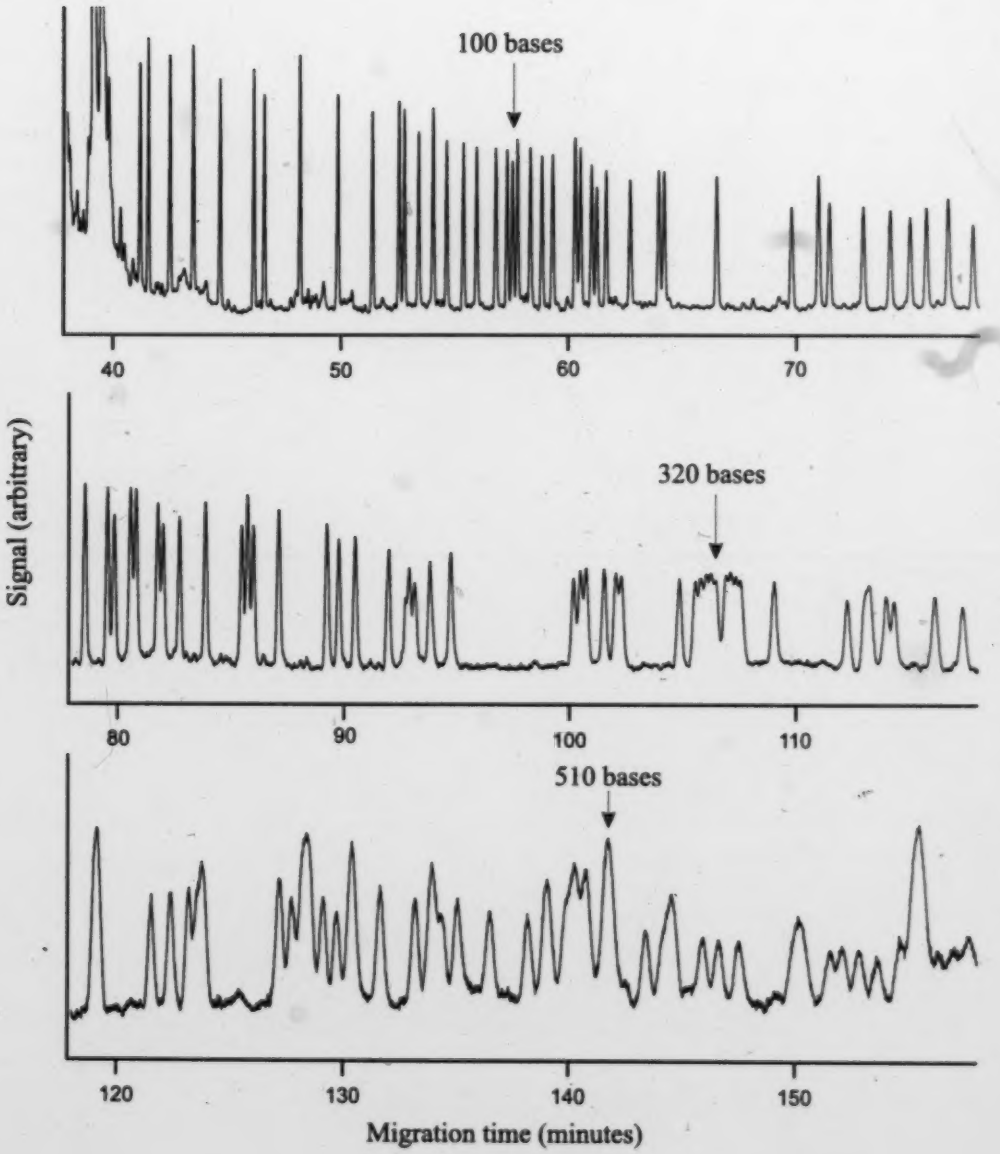


Figure 2.3 at the 100 base triplet. Resolution degrades rapidly for fragments longer than 100 bases in Figure 2.4. This degradation of resolution is indicative of the problem that refilling capillaries has posed.

2.3.2 Single termination sequencing reactions

Actual determination of an unknown DNA sequence requires four-color DNA sequencing chemistry and instrumentation. Each of the A, T, G and C termination reactions uses a primer labeled with a different fluorophore. The sequencing fragments are separated on an instrument that is capable of determining the fluorescence intensity from each individual fluorophore.

For these experiments (and those in chapter 3) I chose to use only the T-termination sequencing fragments for analysis. There are two reasons for doing this. Firstly, the analysis of single termination sequencing electropherograms is much simpler than analysis of four-termination, four-color electropherograms. Four-color sequencing separations require a separate data processing step that converts the spectral information gathered at the detector to an electropherogram trace that represents only one of the four fluorophores. Secondly, the instrument required to perform the separation and detection of the sequencing fragments from a single termination reaction provides a better signal to noise ratio than the four-color instrument. A four-color instrument must divide the time available to collect fluorescence amongst its four spectral channels. This results in poorer signal to noise ratios than are achieved by a single spectral channel instrument. The signal to noise advantage of single spectral channel instruments allowed me to inject very small amounts of DNA on the capillary. Overloading the capillary with DNA may have been possible. The amount of DNA that can be injected onto a polymer filled capillary, and the effect of injecting too much, has not been studied. I used a single-capillary instrument capable of providing high signal to noise ratio to avoid possible overloading.

2.3.3 Failure of refilled capillaries

The most important parameter of a DNA sequencing separation is its read length. Read length is normally defined as the maximum number of bases that an automated base-calling algorithm can correctly determine from a four-color sequencing electropherogram. While many factors influence read length, the most important are

signal to noise ratio and resolution. As mentioned in section 2.3.2, signal to noise ratio was not a factor in these experiments. I chose to define the read length of the single termination sequencing electropherograms as the number of bases at which normalized resolution, calculated by equation 2.1, equals 0.5. Below a resolution of 0.5, individual peaks are not resolved from neighboring peaks when the fragments vary in size by one base. Because it is difficult to determine the widths of peaks that are not baseline resolved, I fit all the peaks in the electropherograms to Gaussian functions with the Peak Fit software package.

Figure 2.5 shows the plot of normalized resolutions versus fragment length, in bases, of the sequencing peaks for each separation of the capillary refilling experiment. Resolution decreases linearly with fragment length. The solid lines are the straight line regression fits to the resolution data from each separation. Resolution becomes considerably worse with each subsequent separation. To determine the read length for each run I simply determined the number of bases for which the regression line equals a resolution of 0.5. Read lengths for each of the four runs in Experiment 1 are reported in table 2.1. From Figure 2.5 it is apparent that resolution for all fragment lengths drops by about 1 unit in the second run on this capillary. Resolution drops another unit for all fragment lengths in the third run on this capillary. By the fourth run, essentially no useful sequencing data is generated.

The number of theoretical plates for each peak in these four runs is plotted versus the fragment length in Figure 2.6. In the first run, the number of theoretical plates increases from a value of about 1 million to a maximum value of about 4 million at 300 bases. The fragment size with the maximum number of theoretical plates decreases dramatically with each successive run on this capillary. Furthermore, the number of theoretical plates at the maximum decreases 2 to 3 fold with each successive run.

The loss in resolution, read lengths, and number of theoretical plates with successive runs could have been due to wider peaks, reduced base spacing as a result of shorter migration times, or a combination thereof. Figure 2.7 is a plot of migration time versus fragment length for all four runs from Experiment 1. Except for the very long fragments at the end of run 1, migration time is a linear function of fragment length.

Figure 2.5 Normalized resolution versus fragment length for runs from experiment 1. Horizontal dashed line corresponds to a resolution of 0.5

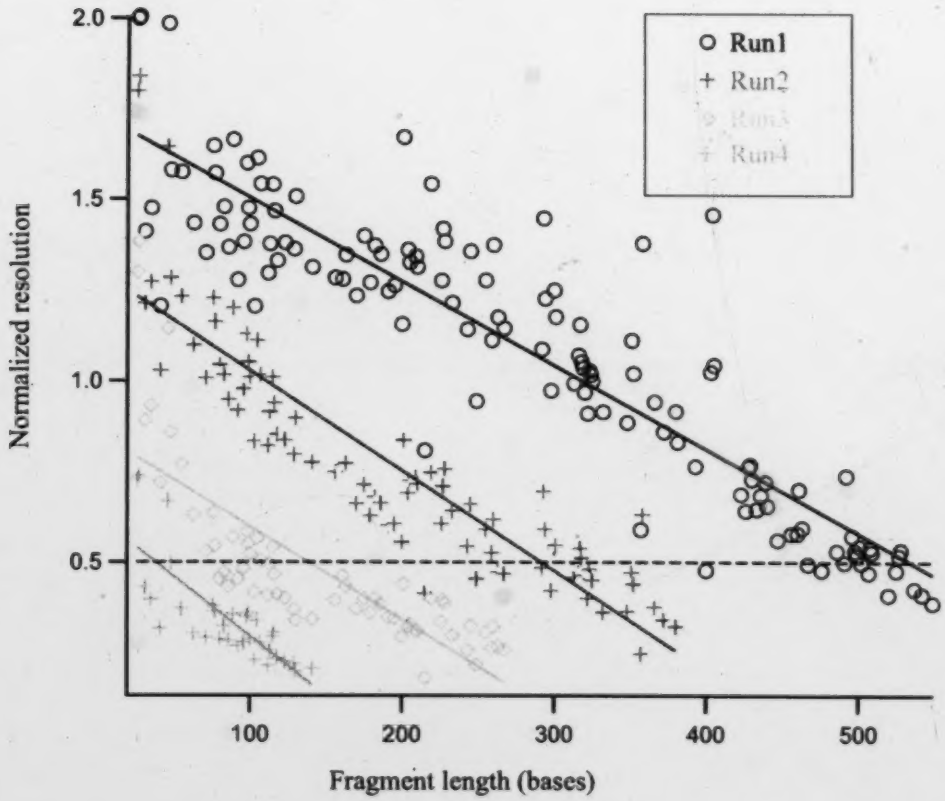


Table 2.1 Read lengths for runs from the capillary refilling experiments (experiment 1)

Run number	Read length (bases)
1	534
2	291
3	139
4	38

Figure 2.6 Number of theoretical plates versus fragment length for experiment 1

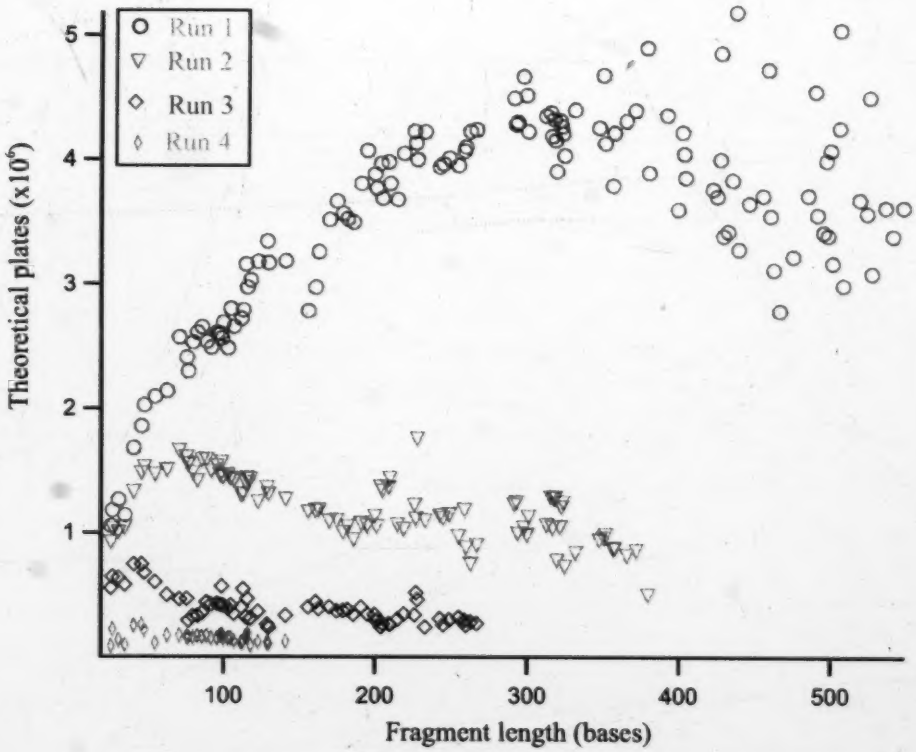
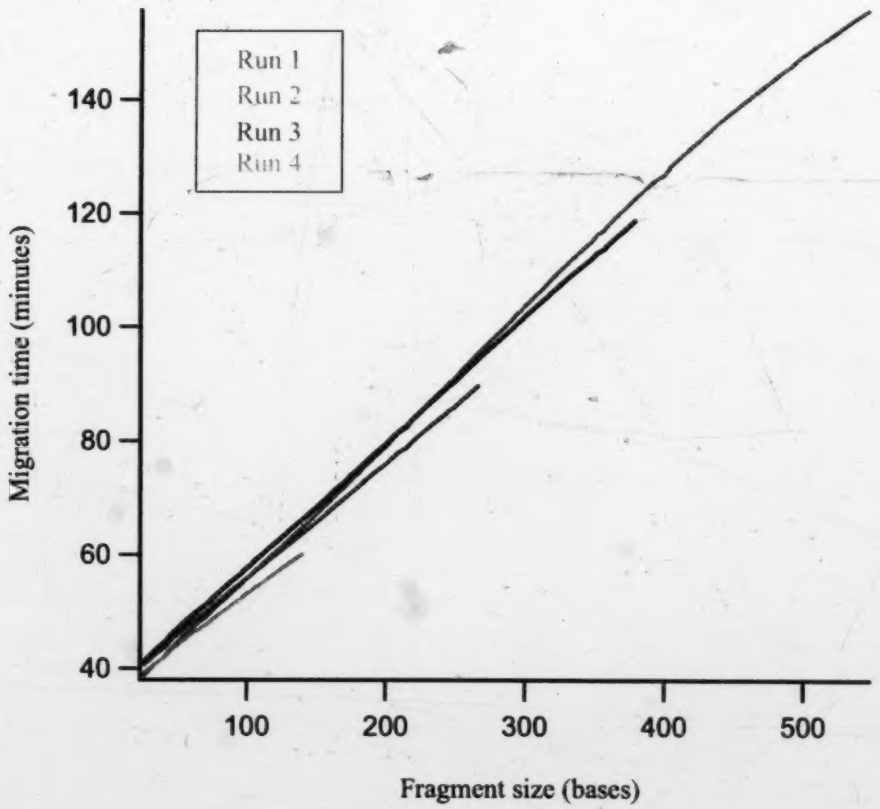


Figure 2.7 Migration time versus number of fragment length for experiment 1



However, with successive runs on this capillary the slope of the plots decreases slightly, as the migration times for a particular fragment decrease slightly with successive runs. The maximum difference in migration times between runs at any fragment size is 4.5%.

Although measurable, the decrease in migration times contributes to only a small fraction of the large drop in resolution and theoretical plates with each successive run. Widening of the peaks is responsible for the majority of the loss in theoretical plates. Figure 2.8 is a plot of peak-width at half-maximum versus fragment length for Experiment 1. Peak widths at all fragment lengths increase by a factor of 1.5 to 2 with each successive run. A 1.5 to 2 fold increase in peak widths will lead to an equivalent reduction in resolution. Theoretical plates are more sensitive to increases in peak widths simply because the value for peak widths is squared. For instance, a doubling of peak width will lead to a four-fold reduction in the number of theoretical plates.

One peculiarity of Figures 2.5 and 2.6 is the considerable scatter to the measured values for resolutions and plate counts. The primary source of this scatter is inconsistent base spacing values. Base spacing is defined as:

$$Basespacing = \frac{\Delta t_m}{\Delta M_D} \qquad (2.3)$$

where t_m is the migration time of an individual peak and M_D is the fragment length for that peak, in bases. Figure 2.9 is a plot of the base spacing versus fragment length for the first run of experiment 1. Notice that base spacing increases slightly until about 300 bases before beginning to decrease. The overall decrease in base spacing occurs as DNA fragments get closer in size to the limiting DNA size, M^*, as described by the biased reptation model (Chapter 1). However, the variation in base spacing from peak to peak is larger than the overall decrease in base spacing. DNA sequencing fragments are not polymers composed of repeating units of identical chemical composition. The primary sequence of the DNA fragments does have some influence on their mobility. These slight differences in mobilities due to DNA sequence are unpredictable, but are consistent for a particular DNA sequence. Variation in base spacing accounts for the majority of scatter in

Figure 2.8 Peak width at half maximum versus fragment length for runs from experiment 1

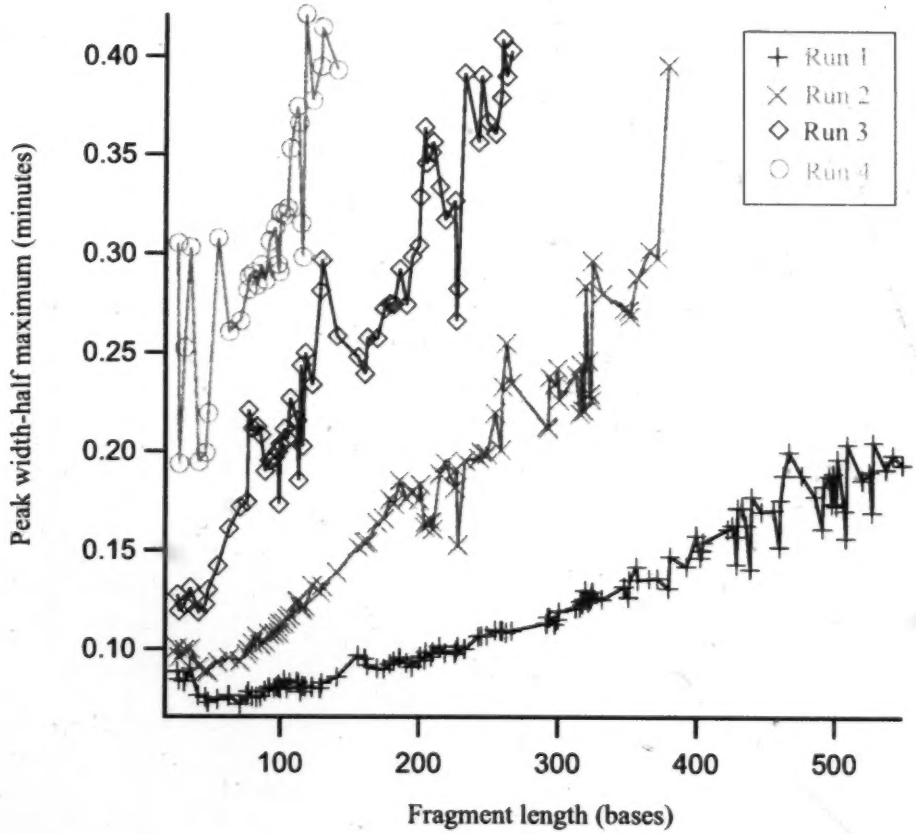
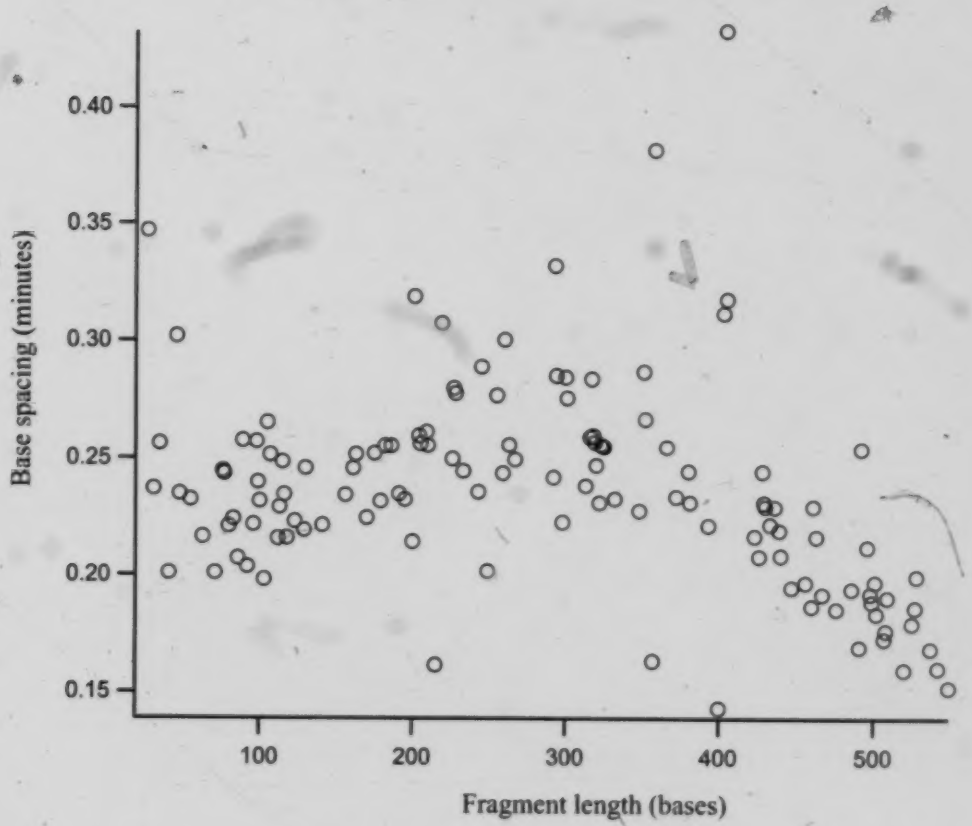


Figure 2.9 Base spacing versus fragment length for the first run from experiment 1



the data in Figures 2.5 and 2.6.

Scatter in the peak width data accounts for a small part of the scatter in the resolution and plate count plots. There is some error introduced into the peak width measurements by the peak fitting procedures. The peak width curve (Figure 2.8) for the first run is fairly smooth for fragment sizes less than 450 bases, after which the points become more scattered. The peak fitting software can make very good fits to peaks that are well resolved. Peaks that have resolution values near or below 0.5 are not as well fit by the software. The same effect is evident for the peak width curve of the second run (Figure 2.8) except that points become more scattered at 250 bases. Because the resolutions of the third and fourth runs are so poor, their peak width plots are much more scattered than the first two runs.

2.3.4 Stability of the Hjerten coating

Widening of the peaks may have been due to an increase in EOF caused by degradation of the capillary coating. The additional EOF could have caused mixing of the zones during the separation and led to peak broadening[18]. I also attribute the decreased migration times seen in Figure 2.7 to failure of the capillary coating. The increase in EOF in a capillary with a failing coating would have been in the direction end of the injection end. While one might expect that this should cause an increase in the migration times, I believe that the effect of the additional EOF was to pump the polymer solution out of the injection end of the capillary. This would have left a plug of buffer which did not contain sieving matrix at the detection end of the capillary. Without sieving matrix, the DNA fragments would have had a higher mobility through the final section of capillary. This may have compensated for EOF in the direction opposite to the mobility of the DNA fragments and might explain the curves seen in Figure 2.7.

Earlier work by Dovichi's group suggests that the capillary coating is stable to chemical degradation up to at least 115 days[19]. Experiment 2 was designed to determine if the failure of the coating was due to alkaline hydrolysis of the silicon-oxygen linkages that bind the coating to the wall. Experiment 2 used capillaries that were aged in a buffer of the same composition as the sieving matrix. If alkaline hydrolysis was indeed responsible for the failure of the coatings, coated capillaries aged in the running buffer

should perform as poorly as a single refilled capillary over time. The resolutions versus number of bases for runs 1, 2, 4 and 5 of experiment 2 are plotted in Figure 2.10. Run 3 failed because an air bubble was introduced into the capillary during filling with sieving matrix. Straight line regression fits are included for each run. Clearly resolution did not degrade with subsequent runs. Read lengths for each run are reported in Table 2.2. Plots of the number of theoretical plates versus fragment length for experiment 2 are shown in Figure 2.11. The maximum of theoretical plates versus fragment length was about 4 million at 350 bases. There is little difference in the plate counts for each of the four runs. These results are similar to those obtained for the first run in Experiment 1. Peaks for the area between 150 and 250 bases in the first run of experiment 2 had lower resolutions and theoretical plates than was expected from the rest of the curve. The peaks in this area of that particular electropherogram tailed slightly, contributing significantly to their width. The source of this tailing remains a mystery, but it fully disappeared by 300 bases. Figure 2.12 is a plot of migration times versus fragment length for the electropherograms in experiment 2. The first three runs have nearly identical migration times; slight differences may be due to slight difference is the capillary lengths. The fourth run has longer migration times. The larger intercept and slope of the curve from the fourth runs resembles a sequencing run done at a lower electric field strength. All runs were done at the same field strength, however, the possibility exists that conditions during the sample injection created a zone of high resistance at the injection tip of the capillary[20]. If this is true it would cause a large voltage drop at the tip of the capillary and would lower the actual field applied to the remaining length of the capillary[21]. Peak widths at half maximum are plotted versus fragment length in Figure 2.13. Again it is clear that there is very little difference between the runs except for the small region of wider peaks in the first run due to tailing. Peak widths were wider for the fourth run; this was presumably due to the longer migration times for this run.

Experiment one suggests that failure of the capillary coating was responsible for the degradation of resolution as the capillary was reused. There was no systematic variation in resolution, theoretical plates, peak widths and migration times for the runs in experiment two, even though the capillaries underwent hydrolysis for the same time as

Figure 2.10 Normalized resolutions versus fragment length for experiment 2. Dashed line corresponds to a resolution of 0.5

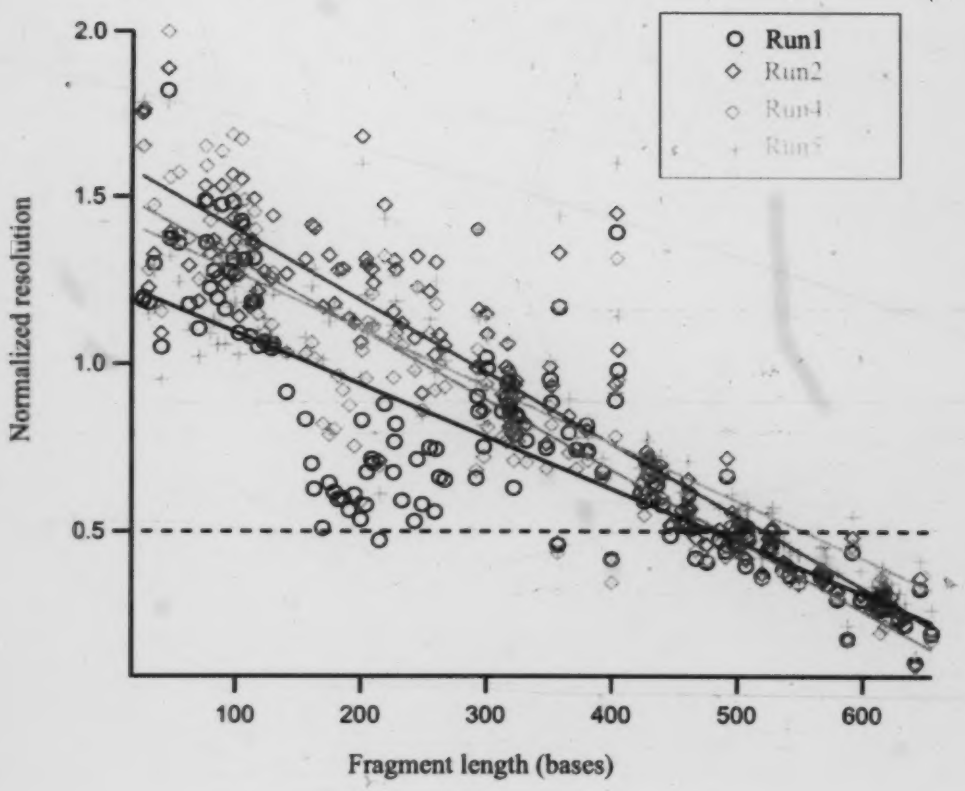


Table 2.2 Read lengths for runs from experiment two

Run number	Read length (bases)
1	480
2	521
4	487
5	552

Figure 2.11 Number of theoretical plates versus fragment length for runs from experiment 2

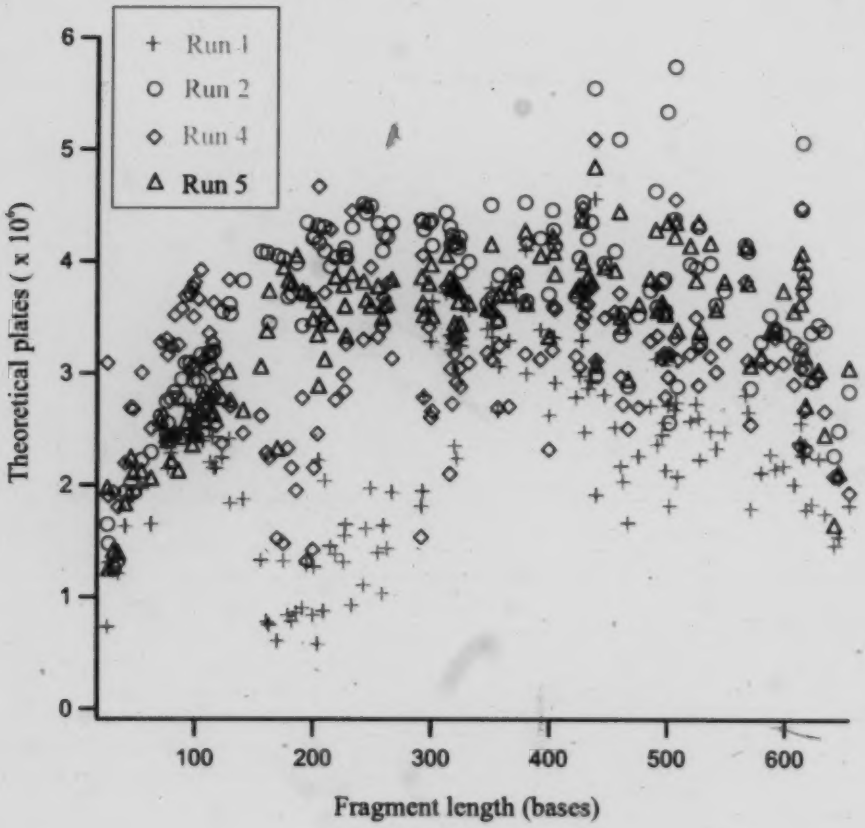


Figure 2.12 Migration time versus fragment length for runs from experiment 2

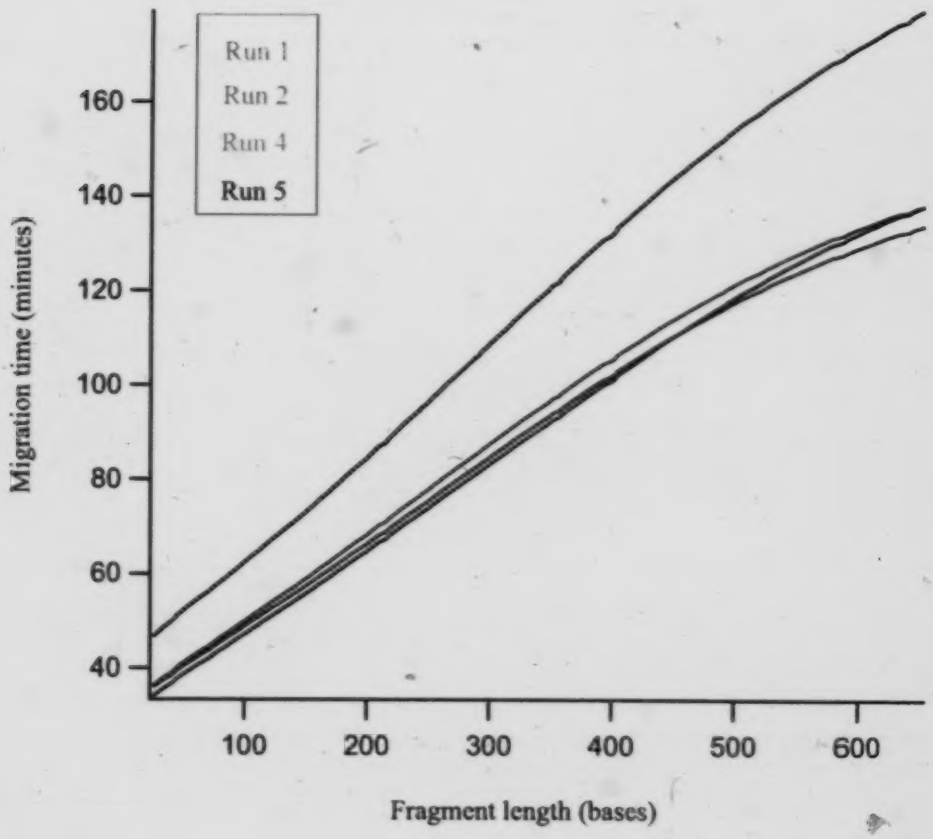
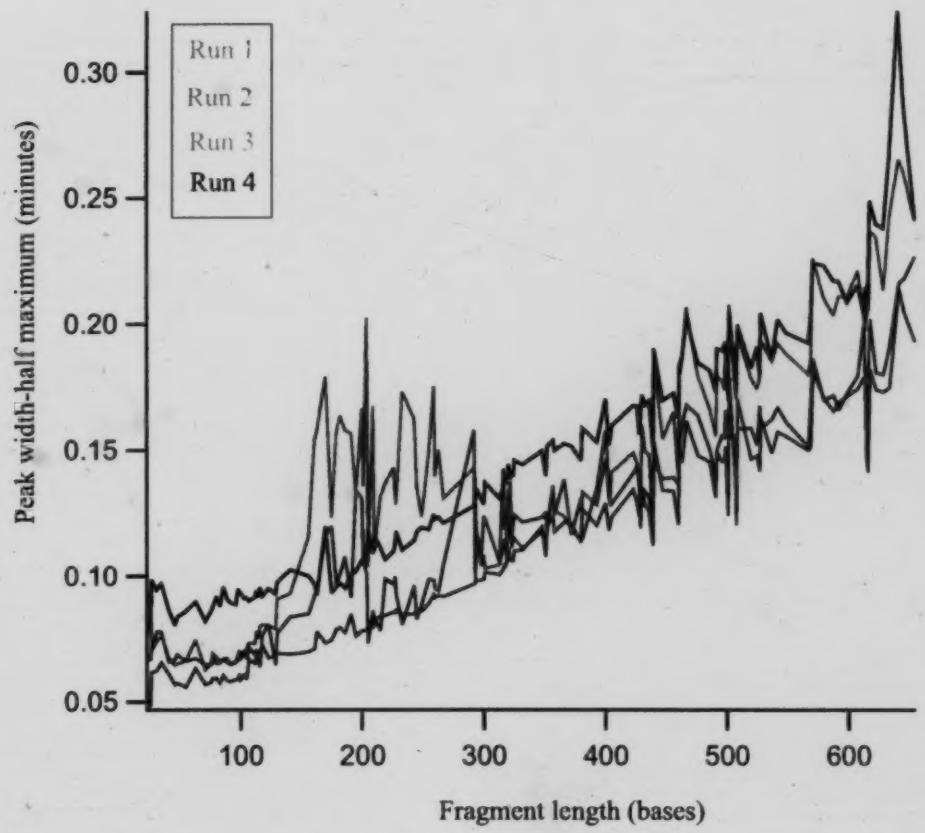


Figure 2.13 Peak width at half maximum versus fragment length for runs from experiment 2



the capillary in the first set of experiments. Alkaline hydrolysis of the coatings does not appear to be a significant factor in coating degradation. I feel that the coating was destroyed by the refilling process itself. High viscosity sieving matrix flowing through the capillary may have sheared the coating off the capillary wall.

2.3.3 Stability of the Grignard-based coating (Novotny coating)

Coatings that are covalently attached to the capillary wall by the Grignard chemistry are reported to be more stable than the Hjerten-style coating[11, 12]. To determine whether the coating based on the Grignard reaction was more resistant to the refilling procedure than the Hjerten coating I used a single Grignard-coated capillary for multiple runs. Read-lengths for 10 sucessive runs are reported in Table 2.3. The average read length obtained for these 10 runs was 482 bases. A plot of normalized resolution versus fragment length for the first and tenth run on this capillary is shown in Figure 2.14. There was little trend towards poorer resolution as the capillary was reused.

This was somewhat of a surprising result. The Grignard coating was designed to be resistant to hydrolysis. The data generated by experiments 1 and 2 suggested that hydrolysis of the Hjerten coating did not play a significant role in the failure of the capillary with successive runs. Some other process, possibly shear forces generated when refilling the capillary with sieving matrix, was likely to blame. In addition to its chemical stability, the Grignard coating appeared to be more physically robust, and able to better withstand the shear forces generated during refilling.

2.3.4 Stability of the poly-DMA sieving matrix

Although I did not do any experiments to determine the shelf-life of the poly-DMA sieving matrix, I did use a single batch of poly-DMA for all three experiments. The quality of run 10 in experiment 3 is similar to the quality of run 1 in experiment 1. The data in Experiment 3 was collected one month after the data for experiment one was collected. Hydrolysis of the poly-DMA polymer does not seem to be of any significance over this length of time.

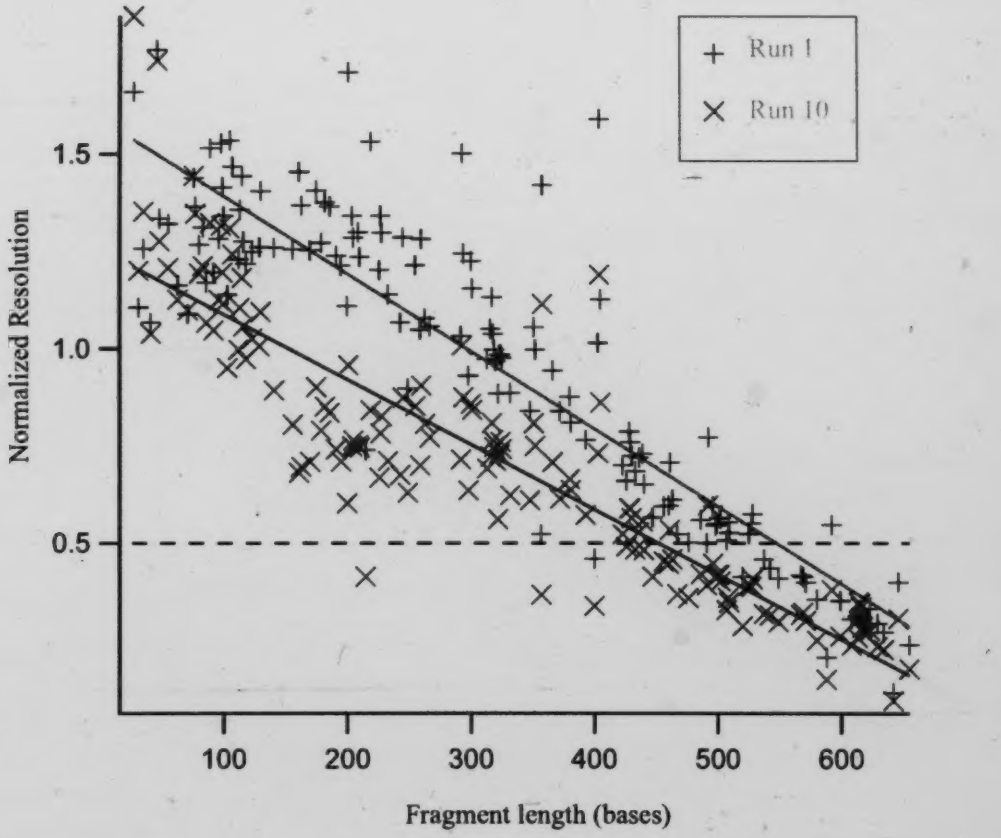
2.3.5 Peak widths and diffusion constants

Peak width data from the first run of experiment 1 (Figure 2.8) was fit to the following equation:

Table 2.3 Read lengths for runs from experiment three

Run number	Read length (bases)
1	547
2	490
3	485
4	473
5	501
6	456
7	450
8	458
9	506
10	452

Figure 2.14 Resolution versus fragment length for run 1 and run 10 of experiment 3. Dashed line indicates a resolution of 0.5



$$W_{1/2} = a + b \times (M_D)^c \qquad (2.4)$$

with coefficients a = 0.0074 b = 2.8×10^{-6} and c = 1.7. This result indicated that peak widths in a good separation scaled as $M_D^{1.7}$. This is in disagreement with the result obtained by Slater and Drouin[22] for slab gels where peak widths were found to scale as $M_D^{1/2}$. My results were in agreement with qualitative differences between sequencing separations done on slab gels and on capillaries. Slab gels tend to have poorer resolutions for the shorter DNA fragments but the resolution drops at a slower rate than in capillary separations. This discrepancy between the results obtained with slab gels and those presented here might be partially explained by the higher field used in CE. According to the biased reptation model (Chapter 1), the diffusion constants will eventually become independent of the size of the DNA molecules as the electric field increased. In this case the peak widths would scale as M_D. This result could indicate that diffusional band broadening is governed by a different mechanism in CE than it is in slab gels. It is also possible that there are non-fundamental sources of band broadening yet to be identified. This would be most fortunate; read lengths for DNA sequencing by CE could be greatly extended if these sources of band broadening, if they exist, were identified and eliminated.

2.4 Conclusion

The ability to replace the sieving matrix is crucial to the success of capillary electrophoresis-based DNA sequencing. I have shown that poly-N,N-dimethylacrylamide can be pumped into and out of capillaries and provide excellent DNA sequencing performance. Unfortunately, the commonly used Hjerten capillary coating is unable to withstand the rigors of replacing the sieving matrix, not because of hydrolysis of the coating but possibly because the refilling process shears the coating from the capillary wall. However, the more robust Grignard-based coating is more resistant to the refilling procedure.

2.5 Bibliography

1. Swerdlow, H.; Harke, H. R.; Wu, S. and Dovichi, N. J. *Journal of Chromatography* **516**, 61-67 (1990).
2. Drossman, H.; Luckey, J. A.; Kostichka, A. J.; D'Cunha, J. and Smith, L. M. *Analytical Chemistry* **62**, 900-903 (1990).
3. Cohen, A. S.; Najarian, D. R. and Karger, B. L. *Journal of Chromatography* **516**, 49-60 (1990).
4. Best, N.; Arriaga, E.; Chen, D. Y. and Dovichi, N. J. *Analytical Chemistry* **66**, 4063-4067 (1994).
5. Ruiz-Martinez, M. C.; Berka, J.; Belenkii, A.; Foret, F.; Miller, A. W. and Karger, B. L. *Analytical Chemistry* **65**, 2851-2858 (1993).
6. Zhang, J. Z.; Fang, Y.; Ren, H. J.; R.Jiang; Roos, P. and Dovichi, N. J. *Analytical Chemistry* **67**, 4589-4593 (1995).
7. Fung, E. N. and Yeung, E. S. *Analytical Chemistry* **67**, 1913-1919 (1995).
8. Menchen, S.; Johnson, B.; Winnik, M. A. and Xu, B. *Electrophoresis* **17**, 1451-1459 (1996).
9. Hjerten, S. J. *Journal of Chromatography* **347**, 191-198 (1985).
10. Gelfi, C.; Desi, P. d.; Alloni, A. and Regetti, P. G. *Journal of Chromatography* **608**, 333-341 (1992).
11. Cobb, K. A.; Dolnik, V. and Novotny, M. *Analytical Chemistry* **62**, 2478-2483 (1990).
12. Engelhardt, H. and Cunat-Walter, M. A. *Journal of Chromatography* **716**, 27-33 (1995).
13. Wu, S. and Dovichi, N. J. *Journal of Chromatography* **480**, 141-145 (1989).
14. Chiari, M.; Nesi, M. and Righetti, P. G. *Electrophoresis* **15**, 616-622 (1994).
15. Gelfi, C.; Simo-Alfonso, E.; Sabastiano, R.; Citterio, A. and Righetti, P. G. *Electrophoresis* **17**, 738-743 (1996).
16. Simo-Alfonso, E.; Gelfi, C.; Sabastiano, R.; Citterio, A. and Righetti, P. G. *Electrophoresis* **17**, 723-731 (1996).
17. Simo-Alfonso, E.; Gelfi, C.; Sabastiano, R.; Citterio, A. and Righetti, P. G.

Electrophoresis **17**, 732-737 (1996).

18. Oefverstedt, L. G.; Johansson, G.; Froeman, G. and Hjerten, S. *Electrophoresis* **2**, 168-173 (1981).
19. Figeys, D. and Dovichi, N. J. *Journal of Chromatograpy A* **717**, 105-111 (1995).
20. Swerdlow, H.; Dew-Jager, K. and Gesteland, R. F. *BioTechniques* **16**, 684-693 (1994).
21. Figeys, D.; Renborg, A. and Dovichi, N. J. *Electrophoresis* **15**, 1512-1517 (1994).
22. Slater, G. W. and Drouin, G. *Electrophoresis* **13**, 574-582 (1992).

CHAPTER 3: THERMAL CONTROL IN CAPILLARY ELECTROPHORESIS DNA SEQUENCING

3.1 Introduction

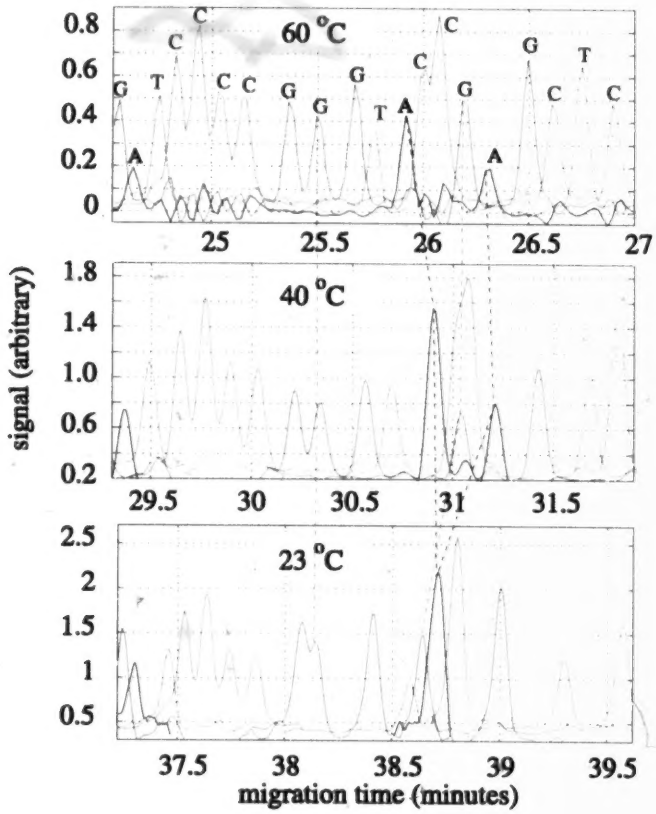
Compressions in DNA sequencing separations are a serious obstacle to accurate DNA sequencing. A compression is caused by a small region of intra-strand secondary structure in a DNA sequencing fragment. This secondary structure increases the mobility of the fragment slightly, so that it co-migrates with DNA fragments one or two bases shorter. Because these co-migrating DNA fragments are not resolved, the DNA sequence in this area is difficult to determine. Figure 3.1 shows a short section of a sequencing electropherogram containing a compression and an electropherogram of the same DNA sequence where the compression has been resolved.

Compressions can be eliminated by increasing the stringency of the denaturation conditions used to keep the DNA single-stranded during a sequencing separation. 7 M urea is the most commonly used denaturant for DNA sequencing. While it is also possible to prepare sequencing gels and sieving matrices containing both urea and formamide as denaturant, the simplest way to eliminate secondary structures in DNA sequencing fragments is to perform the sequencing separation at elevated temperatures[1]. Several studies where capillary electrophoresis DNA sequencing was performed with replaceable sieving matrices at elevated temperatures up to 60 °C have been reported[2, 3]. Performing CE-sequencing separations at elevated temperatures has the added benefit of delaying the onset of biased reptation[4], which limits the read lengths of separations performed at high electric fields[5].

Because of the elimination of compressions and the possibility of extended read-lengths, it is very desirable to perform sequencing separations at elevated temperatures. Surprisingly, no systematic studies have investigated the effect that heating the capillary has on band-broadening. In principle, performing sequencing separations at elevated temperatures is a simple matter of heating the capillary. However, there are several methods of heating the capillary currently being used. Some commercial CE instruments use a liquid jacket around the capillary to heat it. Other instruments heat the air in the space surrounding the capillary. Still other instruments seat the capillary up against a heated plate.

Heating the air-space surrounding the capillary is the simplest way to heat many

Figure 3.1 The effect of temperature on compressions in DNA sequencing electropherograms. 60°C eliminates the compression seen at 23°C.



capillaries in a multi-capillary instrument. Unfortunately, the low heat capacity and high convection of air make precise control over the temperature difficult. The temperature of different points inside near the capillary can vary significantly from the set temperature. Unless the heating system is carefully tuned, the temperature will also oscillate around the set temperature.

To avoid the problems associated with heating the air space around the capillaries, I constructed a solid-state heating unit with proportional-integral feedback control that maintained the set temperature within one-tenth of a degree. To determine the effect of oscillations around the set temperature on sequencing separations I modified the characteristics of the proportional-integral control system to induce oscillations of 0.07°C to 0.75°C about the set temperature and performed DNA sequencing separations at several different oscillation conditions.

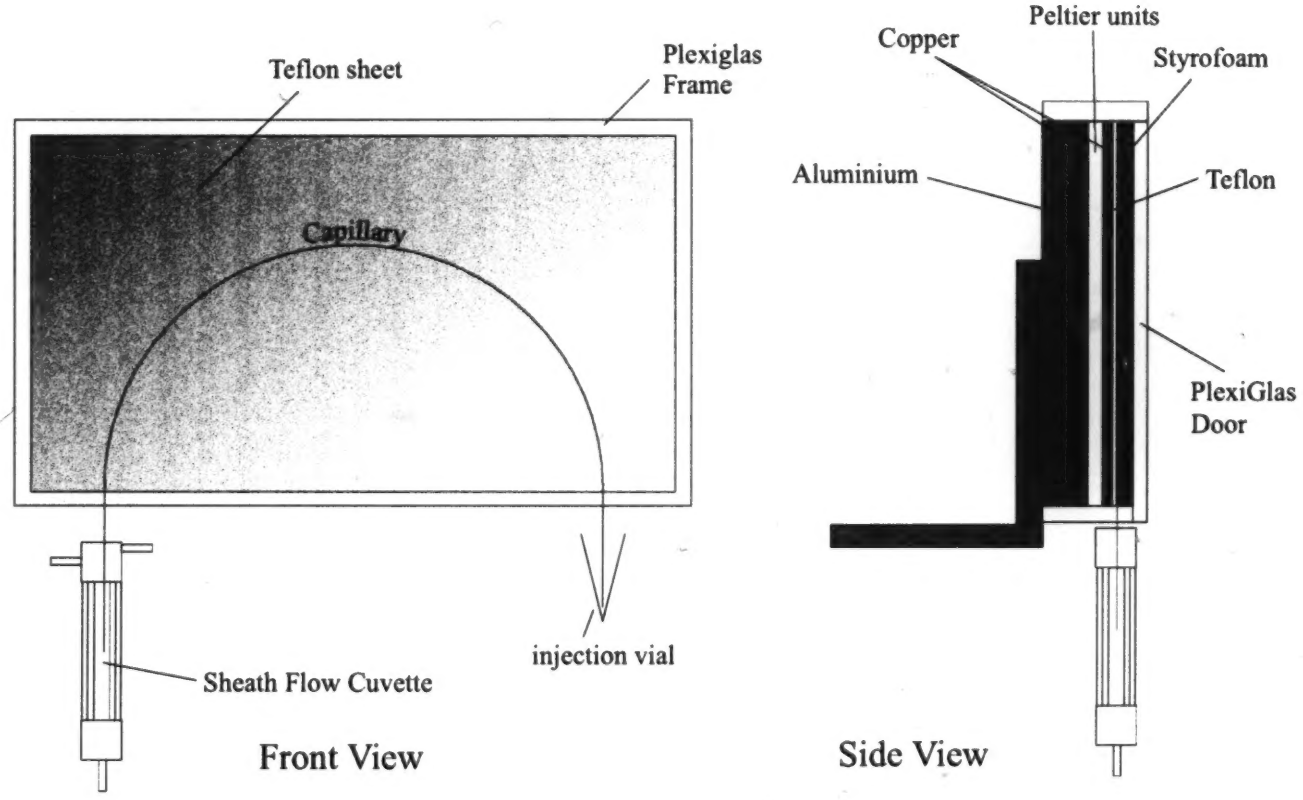
3.2 Experimental

3.2.1 Solid-state temperature controller

A diagram of the temperature controller that we constructed is shown in Figure 3.2. The heating unit consisted of eight 1 cm^2 Peltier units (Melcor, CP 1.4-7.10, 3.9 A, 0.85 V) wired in series and sandwiched between two copper plates 1 mm thick , 22.8 cm long and 14 cm wide. To the outside of one copper plate was glued a 0.5 mm thick Teflon sheet; this was the side against which the capillary rested. The other copper plate was screwed to an aluminum plate of dimensions 6 mm x 24 cm x 15.8 cm. A Plexiglas box with a door at the front enclosed the copper plates, but left the aluminum plate exposed to the air to allow it to radiate excess heat. The entire device was mounted above the sample injection vial and sheath flow cuvette as shown in the left panel of Figure 3.2. The capillary ends were inserted into the sheath flow cuvette and running buffer vial and the capillary placed against the Teflon sheet covering the copper plate. A piece of 1 cm x 22.8 cm x 14 cm Styrofoam was then placed on top of the capillary and the Plexiglas door of the apparatus was closed to firmly set the capillary up against the Teflon sheet.

The copper plates were heated by passing current through the Peltier devices. The current level required precise control to maintain a constant temperature in an external environment that included changing room temperatures due to air conditioning, drafts and

Figure 3.2 Heating unit used to DNA sequencing at elevated temperatures



breezes. Control was provided by an analog proportional-integral (PI) feedback controller circuit constructed in-house. Figure 3.3 contains the schematic of the control circuit. The diagram is divided into three regions (enclosed by boxes) for simplicity. The portion enclosed by the top box converted the frequency of a digital waveform from the National Instruments MIO-16X data acquisition card to a voltage. This was necessary because the data acquisition card did not have enough analog outputs available to control the CE instrument and the heating unit. The output voltage of this portion of the circuit controlled the set point, or target temperature of the heating device. The lower left panel of this circuit was the actual PI controller unit. It provided a voltage proportional to the difference between the resistance of a reference resistor and a 10 kΩ thermister that was placed in contact with the Teflon sheet beside the capillary. This portion of the circuit outputs either a positive voltage (if the thermister was cooler than the set point) or a negative voltage (if the thermister was warmer than the set point). The lower right panel of the circuit provided current to the Peltier units. This current is proportional to the output voltage from the PI section of the controller and was positive (which heated the copper plate in contact with the capillary) or negative (which cooled the same plate). The lower panels of this circuit contained three potentiostats which were used to empirically tune the controller to lock in properly at 45 °C with no oscillations.

3.2.2 Temperature measurement

The temperature of the heating unit was monitored and recorded by computer during all separations. The temperature measurements were made by comparing the resistance of a second 10 kΩ thermister that was placed alongside the capillary with a reference resistor. The circuit diagram for the temperature probe is shown in Figure 3.4. The output voltage from this circuit was monitored by the MIO-16X data acquisition card and recorded along with the sequencing electropherograms by a Macintosh computer. I did not determine the accuracy of this temperature measurement system, however I always ensured that the temperature was within one degree of the temperature measured with a commercial thermocouple thermometer. Noise in the measured temperature had a standard deviation of 0.005 °C. Because of this noise, only temperature changes of 0.01 °C or larger could be measured.

Figure 3.3 PI controller circuit

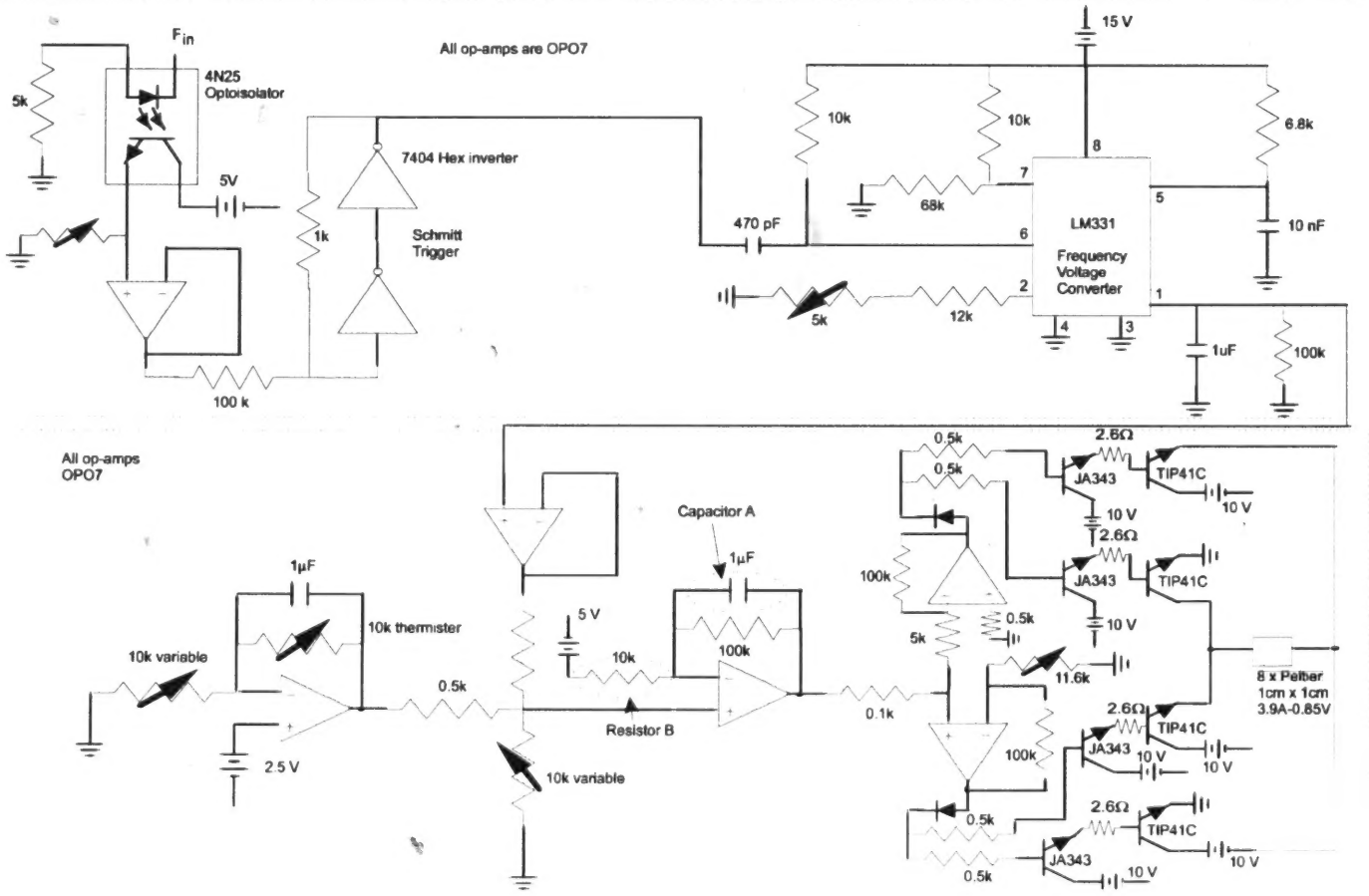
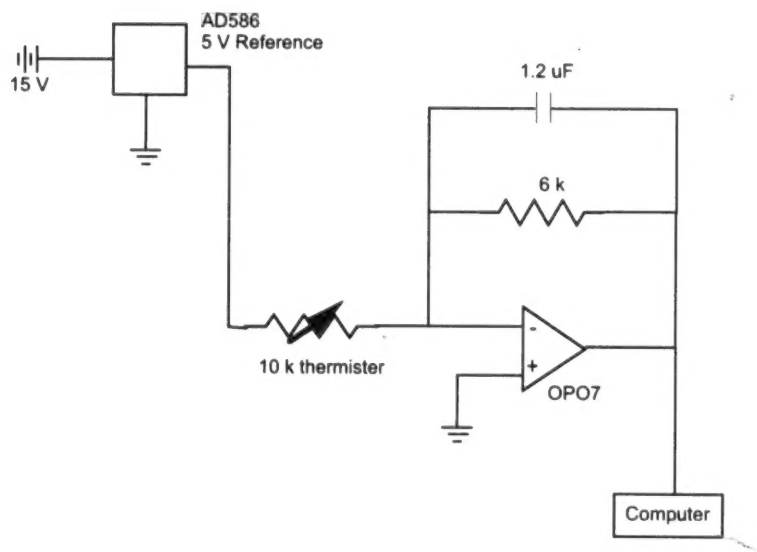


Figure 3.4 Temperature monitor circuit diagram



3.2.3 Preparation of T-terminated sequencing samples

T-terminated sequencing samples were prepared by cycle sequencing. 1.5 μL of M13mp18 single-stranded DNA (Amersham), 2 μL of M13-21 Tamra-labeled primer (Perkin-Elmer), 4 μL of Thermosequenase™ T reagent (Amersham) and 8.5 μL of water were mixed in a 200 μL PCR tube. Thermal cycling was done with a MJ Research PTC-100 thermocycler with heated lid. Cycling conditions were as follows: 30 cycles of 95°C for 30 seconds and 55°C for 30 seconds. After cycling the samples were purified by a Microcon 30 spin tube. Samples were eluted from the Microcon tubes in 15 μL of deionized formamide, which were then ready for injection on the capillaries.

3.2.4 Sieving matrix

The 6% poly-N,N-dimethylacrylamide sieving matrix used for all the sequencing runs was provided by the Perkin-Elmer corporation. It is a highly purified poly-DMA matrix that adsorbs to the capillary wall, which reduces electroosmotic flow and eliminates the requirement for the capillary to be coated[6]. The buffer used was 100 mM Tris, 100 mM TAPS at pH 8.0. Viscosity of this sieving matrix was not measured but is approximately 1500 cp, low enough to be pushed through the capillary with a syringe.

3.2.5 Sequencing separations at a constant temperature.

A single capillary, 60 cm long, was placed in a 5 capillary electrophoresis instrument[7]. The instrument was configured to use the 543 nm line of the helium-neon laser for excitation and 580DF10 bandpass filter, GRIN lens, and single photon counting module for detection. The fluorescence intensity was sampled at 4Hz. The capillary was filled with sieving matrix using a device similar to the one described in Figure 2.4. The only difference was that the pressure was provided via a 3 mL disposable syringe and syringe pump (Razel Scientific) instead of an HPLC pump. The capillaries were refilled for 45 minutes between runs; this ensured that all the polymer was replaced. An empty capillary could be filled in less than 10 minutes with this system. After refilling was complete, the injection tip of the capillary was removed from the refilling apparatus and placed in a 400 μL vial of running buffer (100 mM Tris, 100 mM TAPS, pH 8.0). The heating unit was engaged and set to 45 ± 1.0°C. 9000 Volts was applied across the capillary in reverse polarity (cathode at detection end) for approximately 2 minutes. If the

current was stable the polarity was reversed and 9000 V was applied to the capillary for 10 minutes before injection. An unstable current at this stage indicated that an air bubble was introduced into the capillary during the refilling process. Capillaries with unstable currents were refilled. Samples were injected for 20 seconds at 6000 V. After injection the sample was replaced with running buffer and the separation continued at 9000 V.

3.2.6 Sequencing separations with an AC temperature fluctuation.

To perform sequencing separations at an elevated temperature with a 0.1 °C to 0.7°C AC component to the temperature profile required modifying the characteristics of the PI feedback temperature controller. I removed the capacitor marked A in Figure 3.3 and lowered the value of resistor marked B. Resistor B was set between 1300 Ω and 1600 Ω, depending on the magnitude of temperature fluctuation required. When setting the desired temperature fluctuations the actual temperature was monitored while the gain was adjusted. These changes caused the temperature of the heating unit to oscillate around the 45°C set point. Other than the described changes to the heating system controller, all of the parameters for the separations were exactly as described in section 3.2.5.

3.2.7 Data analysis

Electropherograms were imported into PeakFit v 4.0 (Jandel Scientific) and Gaussian curves fit to each peak. Curves were only fit to peaks that were recognizable as belonging to a certain fragment length(s). Curves were not fit to peaks later in the runs with very poor resolution because it was difficult to distinguish individual peaks or groups of peaks. Normalized resolution was calculated according to equation 2.1. Calculation of the number of theoretical plates was according to equation 2.2. Temperature fluctuation (ΔT) was defined as the peak to peak amplitude of the AC component of the measured temperature data. The average temperature fluctuation was estimated by measuring the temperature fluctuation at regular intervals and averaging these measurements.

3.3 Results and Discussion

3.2.1 PID controller theory

Feedback control theory of dynamic systems is a science unto itself. The following is a qualitative explanation of PI control. Analytical analysis of dynamic systems and

control theory are included in engineering textbooks on the subject[8]. The purpose of the classical PID (proportional-integral-derivative) controller is to maintain a system (in our case the capillary heating block) within a very narrow range of a steady state set point (the desired temperature of the block) in the face of changing internal and external conditions. The most important parameter of a PID type control system is its gain. This gain determines the magnitude of the response the controller applies to correct for a difference in the set point temperature and the measured temperature. In our system this gain is provided by the amplifier portion of the circuit shown in the bottom-left box of Figure 3.3. The gain of the second amplifier (containing Resistor A and Capacitor B) controls the amount of current that is supplied to the Peltier units. It also determines the width of the proportional band. At temperatures within the proportional band the controller will supply an amount of current that is proportional to the difference between the actual temperature and set temperature. If this gain is too low, and the proportional band is too wide, the controller will not supply enough current to make up for the loss of heat from the system and the heating plate will never reach its set point. If this gain is too high, and the proportional band is too narrow, the controller will supply more current than is needed to make up for heat loss and the temperature will overshoot the set point. At this point the controller will reverse the current in attempt to cool the plate and the temperature will drop below the set point. The temperature will continue to oscillate around the set point with an amplitude dependent on the gain. Figure 3.5 demonstrates the effect of too much and too little gain. The actual size of the proportional band depends on the physical characteristics of the system such as the heat capacities and conductivities of the copper plate, the distance between the Peltiers and the thermister, and the amount of heat lost from the copper plates to the environment. By using too much gain I was able to induce small temperature fluctuations in the copper plate. The measured temperature profiles for three different values of resistor B are shown in Figure 3.6. For clarity, the traces are offset in their DC value. The magnitudes of the oscillations were not altered. While the amplitude of the temperature waveform increased with increased gain, the period of the waveform did not change dramatically. The frequency of this oscillation is known as the characteristic frequency. The characteristic frequency is an accurate representation of the

Figure 3.5 Characteristic temperature profiles for a proportional feedback controller with too much gain (top) and too little gain (bottom)

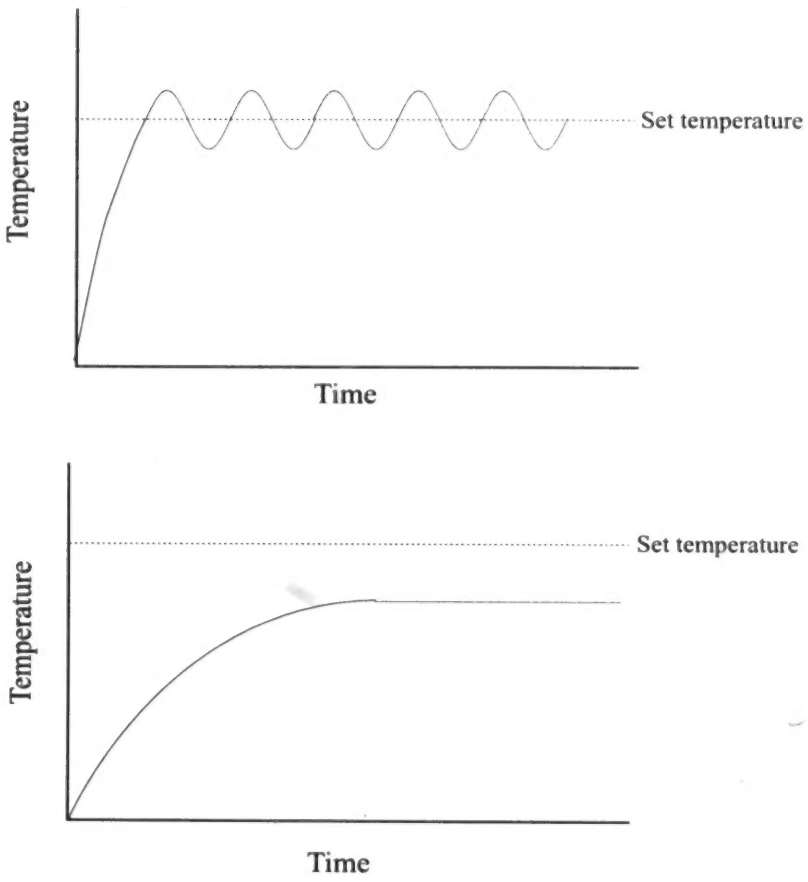
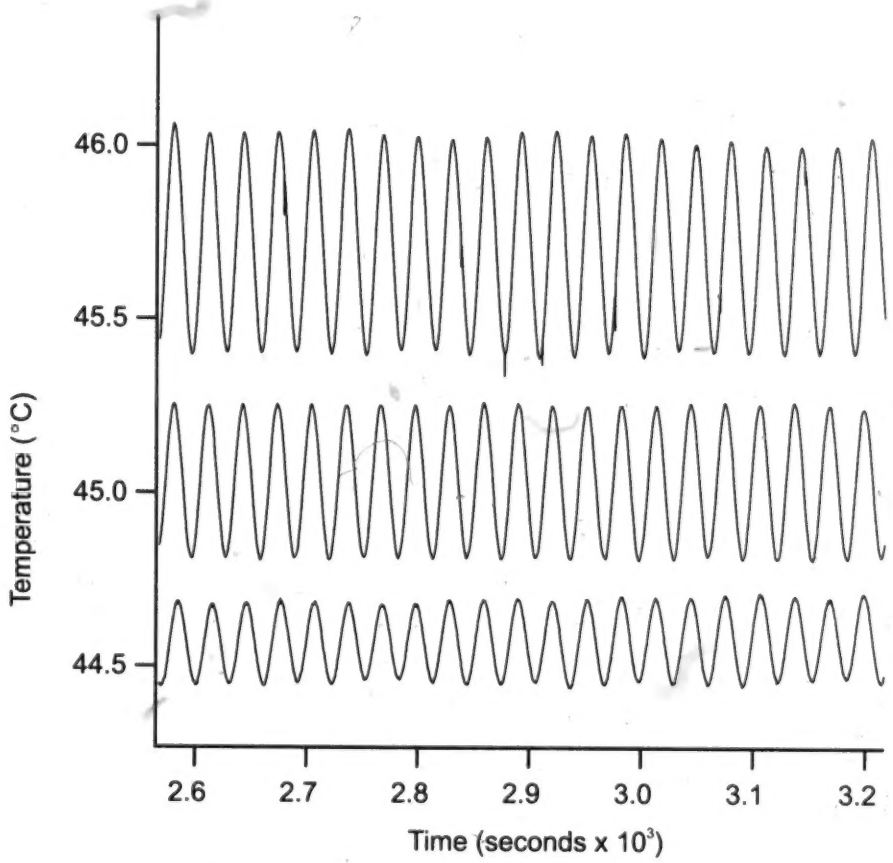


Figure 3.6 Temperature oscillations at different gains. The traces are temperature profiles for different values of resistor B (Figure 3.3). Values of resistor B are: 1300 ohms (top trace), 1400 ohms (middle trace) and 1600 ohms (bottom trace). For clarity the traces are offset in their DC values



responsiveness of the system. The temperature fluctuation and characteristic frequency for each value of Resistor B I used is reported in Table 3.1.

A controller that is able to lock-in at exactly the set temperature requires integral action as well as proportional action. In my case, the integral action is provided by Capacitor A. Integral action adds time dependence to the controller response. It slows the response of the proportional control and gives the system time to settle thermally before the controller attempts to correct for the difference in measured temperature and the set temperature. The integral action prevents oscillations around the set point under high gain conditions, as shown in the top panel of Figure 3.7. It also increases the magnitude of the response to a difference between the actual and set temperatures as time increases, eliminating steady state errors between these two temperatures under low gain conditions, as shown in the bottom panel of Figure 3.7. Ideally, the time constant of integration should match the time constant of equilibration of temperature across the copper plate. Figure 3.7 details the difference between proportional control and PI control of temperature. When tuning a PI controller, the gain should be set until the system just starts to oscillate around the set point, the characteristic frequency is measured and then the gain decreased slightly before adding the integral action. The time constant for the integral and derivative components of the control are derived from the characteristic frequency[9]. Too long of a time constant for the integrator and the controller reacts sluggishly to a sudden change in the temperature of the system.

We did not use any derivative response in our controller (hence PI instead of PID). Derivative response is necessary to ensure the controller can react quickly to a sudden change in temperature. The derivative response is difficult to tune properly and is not needed when the only disturbances that the temperature controller encounters are slow changes in room temperature.

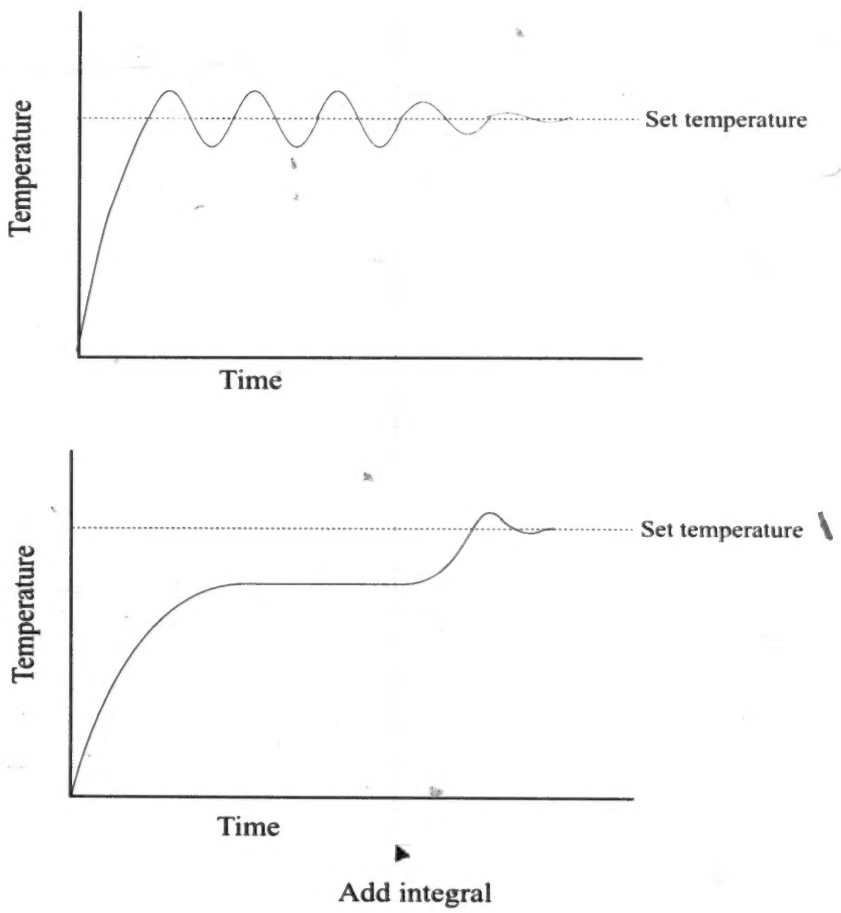
3.3.2 The effect of temperature fluctuations on sequencing separations

The amplitude of the temperature fluctuations tended to vary sightly during a run, usually less than 0.05 °C. Runs where the amplitude of the temperature fluctuation changed more than 0.1 °C during the course of the run were discarded. Generally, larger gains for the PI controller meant larger temperature fluctuations. However, this was not

Table 3.1 Temperature fluctuations and characteristic frequencies for differing PI controller gains.

Resistor B (Ω)	Temperature Flux. (°C)	Characteristic Frequency (Hz)
9000	<0.01	N/A
1450	0.07	0.034
1600	0.27	0.032
1500	0.28	0.032
1600	0.32	0.032
1400	0.50	0.032
1300	0.75	0.032

Figure 3.7 Effect of integral action on temperature profiles for a proportional feedback controller with too much gain (top) and too little gain (bottom)



always the case. For unknown reasons the amplitude of the temperature fluctuation at a particular gain changed from day to day. Although this was not a problem because I constantly monitored the temperature and empirically adjusted the gain to get the desired fluctuation, the variation in temperature fluctuations does highlight the sensitivity of this control system.

Figure 3.8 shows the effect that different temperature fluctuations had on the resolutions of sequencing separations. Three runs were done that had temperature fluctuation in the 0.25 - 0.35 °C range. Of these three runs, only the data from the run with a temperature fluctuation of 0.28 °C was plotted on this graph and all following graphs. This was done to improve the clarity of the graphs. Points from the other runs in this temperature range fall almost exactly on top of points from the 0.28°C run. The linear regression lines that were fit to the data sets are also shown on Figure 3.8. From the regression lines, read length was estimated as the fragment length corresponding to a resolution value of 0.5 (dashed line). Read lengths and average temperature fluctuation are reported in Table 3.2. The trend was to worse resolution and shorter read lengths as the magnitude of the temperature fluctuation was increased. There was very little loss in resolution of the run done with 0.07 °C temperature fluctuation. There was also not much loss in resolution for the run with 0.28 °C temperature fluctuation at fragment lengths smaller than 250 bases. Most of the resolution was lost for the longer fragments. A temperature fluctuation of 0.5 °C caused a loss of about 0.5 in the resolution values at all fragment lengths. No useable sequence data was obtained with a temperature fluctuation of 0.75 °C.

Just how much temperature fluctuation can be tolerated in DNA sequencing applications? Figure 3.9 is a plot of read length versus temperature fluctuation. A straight line regression fit to these points is also included. This plot suggests that any temperature fluctuation will reduce read lengths. This is probably true, however, a temperature fluctuation less than 0.1 °C would probably have an undetectable effect of read length in most cases. According to the regression line, an increase in the temperature fluctuation of 0.2°C results in the loss of 200 bases of sequence. A temperature fluctuation greater than 1°C will destroy any separation.

Figure 3.8 Resolutions versus fragment length for sequencing separations done with different temperature fluctuations. Straight lines are linear regression fits to the data sets. Dashed line indicates a resolution of 0.5

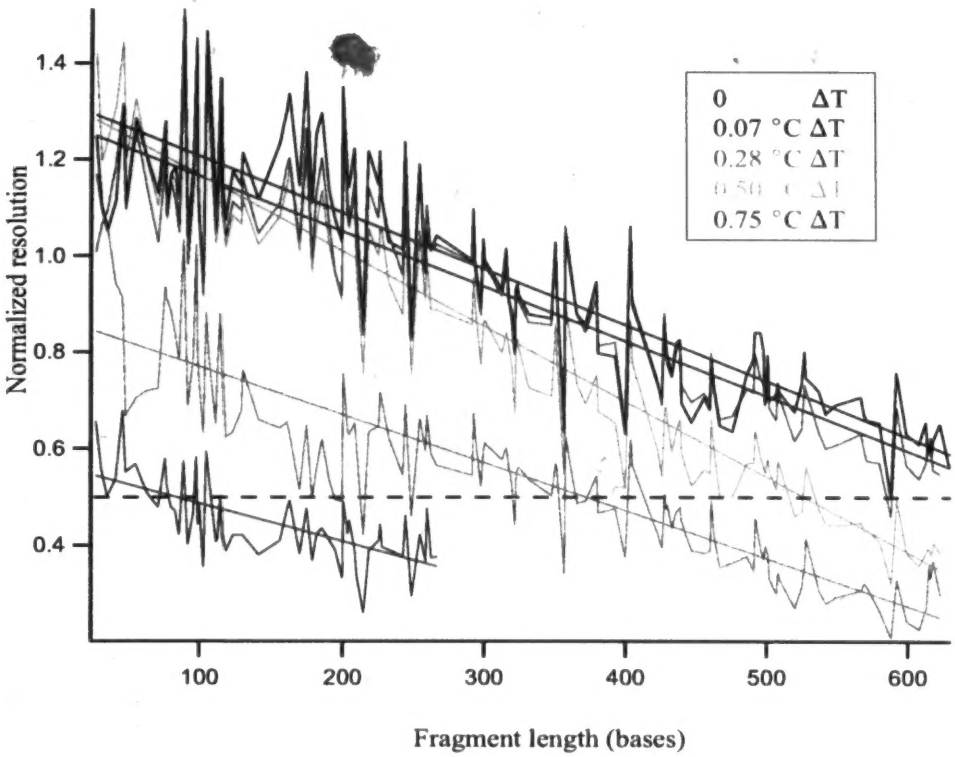
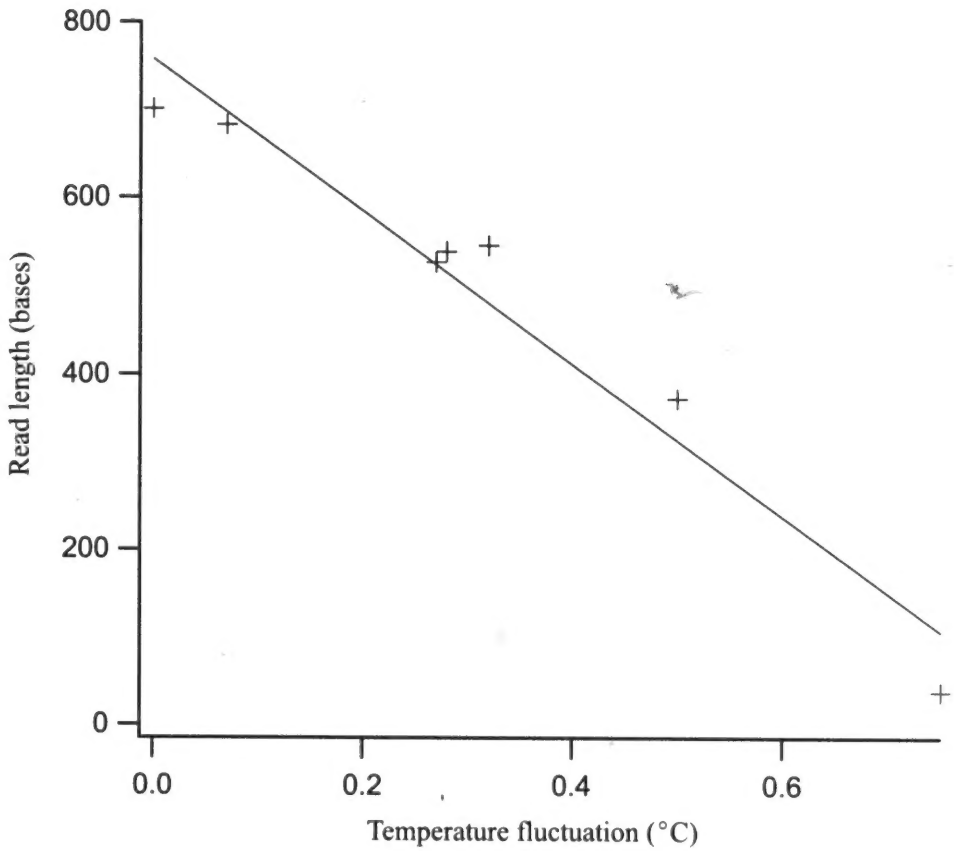


Table 3.2 Read lengths at different temperature fluctuations.

Temperature Fluctuation (°C)	Read length (bases)
None	701
0.07	683
0.27	526
0.28	538
0.32	545
0.50	371
0.75	38

Figure 3.9 Read length versus temperature fluctuation



The number of theoretical plates versus fragment length is plotted in Figure 3.10. The effect of temperature fluctuations are reminiscent of the effect of coating failure described in Chapter 2. Both the fragment length at which the maximum number of plates occurred and the actual plate counts dropped as the magnitude of the temperature fluctuation increased. The run with 0.07 °C temperature fluctuation had similar theoretical plate counts to the run with no temperature fluctuation. Both of these runs reached their maximum number of theoretical plates at approximately 500 bases, after which they decreased slightly. In contrast, the theoretical plates for the run with 0.28 °C temperature fluctuation maintained values close to those for the better runs until 350 bases. At this point the number of theoretical plates dropped noticeably. There was a large drop in the number of theoretical plates for the runs with 0.5 °C and 0.75 °C temperature fluctuations.

Figure 3.11 plots migration time versus fragment length. The migration times for the sequencing fragments changed very little between runs; the slight offsets between the curves may have been due to slight differences in the average temperature. Although the average temperature was between set between 44 °C and 46 °C I did not attempt to set it exactly at 45 °C for every run. Both the slope and intercepts of the migration time versus fragment length curves vary with temperature[4].

The migration time curves indicated that the cause of the decreased plate counts, resolution, and read lengths was peak widening. The plot of peak widths at half maximum versus fragment length (Figure 3.12) confirms this observation.

Figures 3.8 and 3.9 exhibit considerable scatter in the data. As discussed in Section 2.3.3, the majority of this scatter was due to non-uniform peak spacing. The resolution data in Figure 3.8 demonstrated that the changes in peak spacing values were consistent for particular fragment sizes. A data point that was above the regression line in one run was also above the regression line in all of the other runs. Scatter in the peak width data (Figure 3.10) was mostly due to errors in fitting peaks that were not fully resolved (Section 2.3.3). Again, scatter in the peak width data was larger later in the runs and larger for the runs with worse resolution.

The detrimental effect of oscillatory temperature fluctuations on electrophoretic

Figure 3.10 Number of theoretical plates versus fragment length for different temperature fluctuations

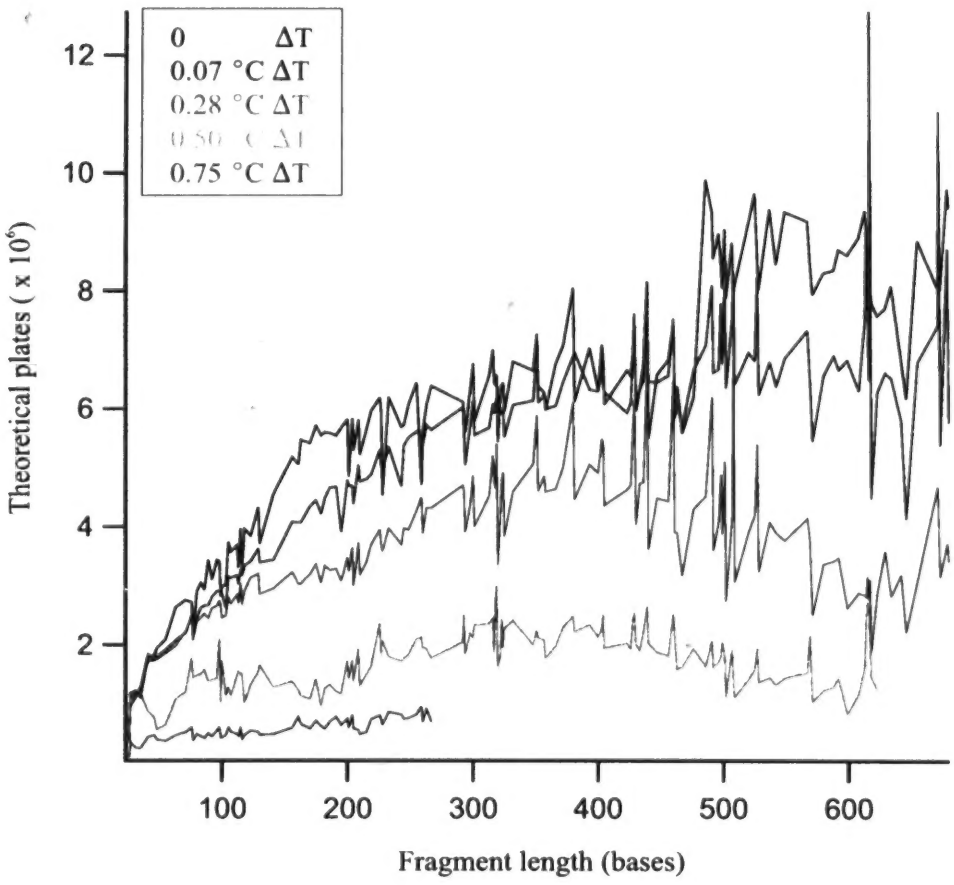


Figure 3.11 Migration time versus fragment length for different temperature fluctuations

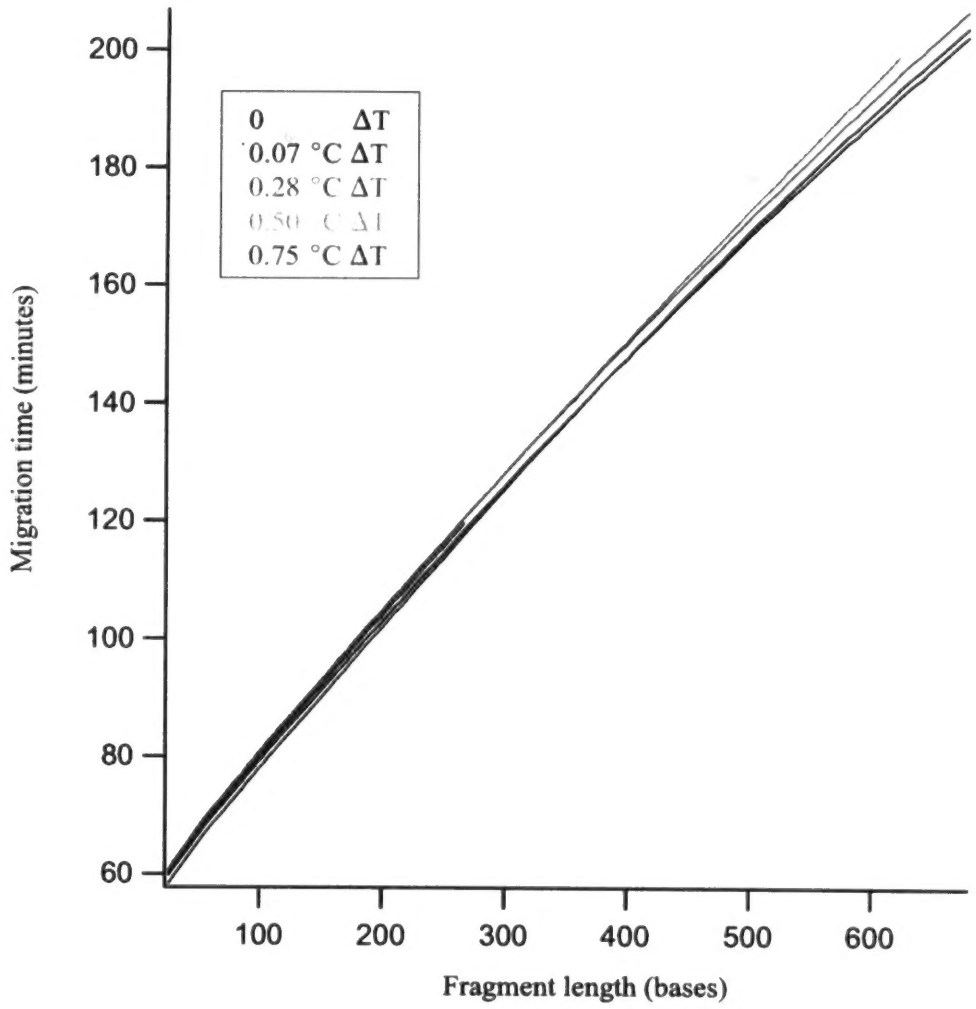
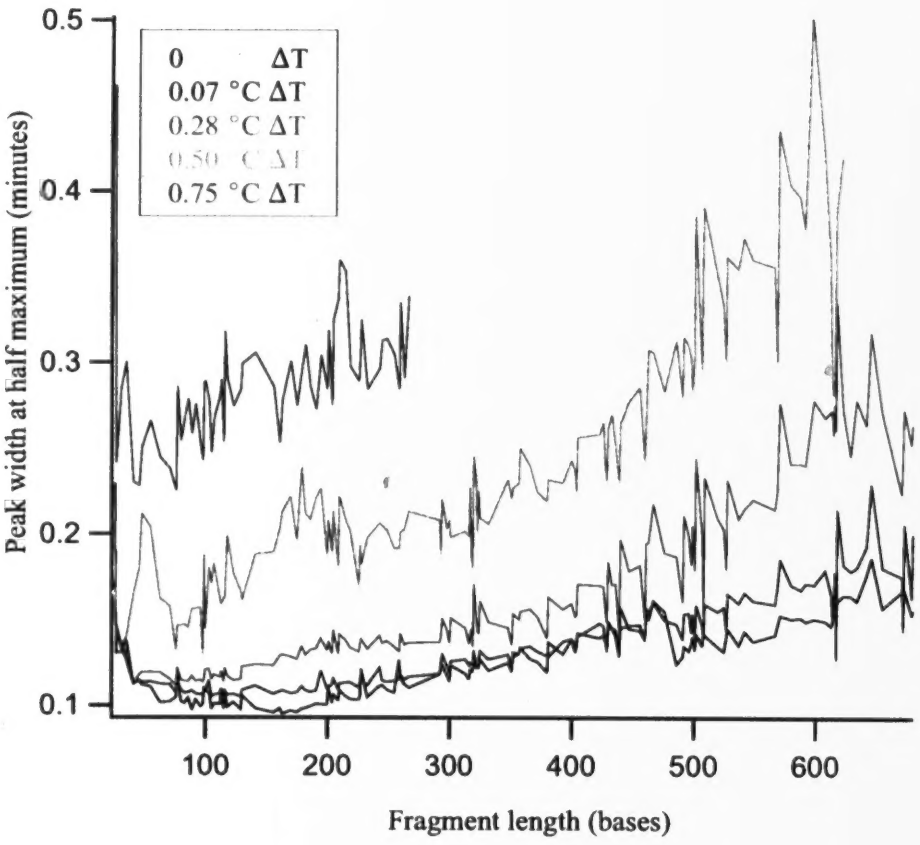


Figure 3.12 Peak width at half maximum versus fragment length for different temperature fluctuations

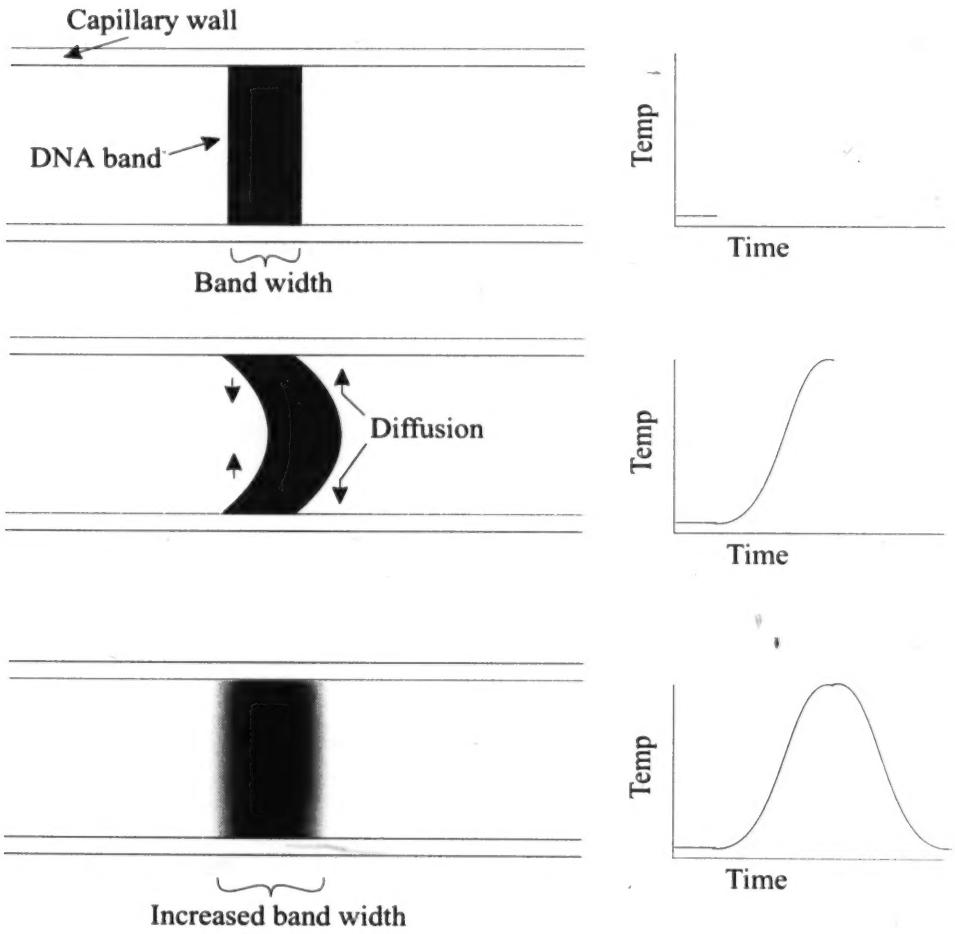


separations has not been described before. I speculate that the band broadening is due to the expansion and contraction of the sieving matrix as the temperature fluctuates. If the sieving matrix was water, which has a density of 0.9902132 $g \cdot mL^{-1}$ and a density change of -0.0004183 $g \cdot mL^{-1} \cdot °C^{-1}$ at 45 °C, an increase of 1°C would cause a 0.04% increase in the volume of the water inside the capillary, which would extrude a 250 μm plug of water from a 60 cm capillary. A temperature change would induce a small flow in the bulk liquid in the capillary. It is possible that this flow had a parabolic profile which would influence the shape of the DNA bands. Although the bands will return back to their original position when the temperature cools, radial diffusion of the DNA molecules while the bands are parabolic would lead to band broadening. This idea is summarized in Figure 3.13. This mechanism is speculative. However, if temperature fluctuation does cause a parabolic flow, modeling has shown that oscillatory parabolic band shapes will indeed lead to band broadening[10].

3.4 Conclusion

I have shown that DNA sequencing by capillary electrophoresis with low viscosity polymer sieving matrix was sensitive to periodic fluctuations in the temperature. To obtain maximum read-lengths sequencing separations required that the periodic temperature fluctuations be minimized. I constructed a solid state heating unit with proportional-integral control that had no detectable temperature fluctuation and a drift of less than 0.1 °C. With this heating unit and a replaceable sieving matrix I could achieve read lengths of greater than 700 bases.

Figure 3.13 Model for temperature fluctuation induced band broadening



3.5 Bibliography

1. Konrad, K. D. and Pentoney, S. L. *Electrophoresis* **14**, 502-508 (1993).
2. Zhang, J. Z.; Fang, Y.; Ren, H. J.; R.Jiang; Roos, P. and Dovichi, N. J. *Analytical Chemistry* **67**, 4589-4593 (1995).
3. Kleparnik, K.; Foret, F.; Berka, J.; Goetzinger, W.; Millar, A. W. and Karger, B. L. *Electrophoresis* **17**, 1860-1866 (1996).
4. Fang, Y.; Zhang, J. Z.; Hou, J. Y.; Lu, H. and Dovichi, N. J. *Electrophoresis* **17**, 1436-1442 (1996).
5. Lu, H.; Arriaga, E.; Chen, D. Y.; Figeys, D. and Dovichi, N. J. *Journal of Chromatography* **680**, 503-510 (1994).
6. Madabhushi, R., S.; Menchen, S., M.; Efcavitch, J. W.; Grossman, P., D *United States Patent* No. 5552028 (1996).
7. Zhang, J. in *The Sheath Flow Cuvette in High-Sensitivity Fluorescence Detection for DNA Sequencing in Single and Multiple Capillary Systems,* Ph.D. thesis, Department of Chemistry, University of Alberta, Edmonton, 1994.
8. Franklin, G. F.; Powel, J. D. and Emami-Naeini, A. *Feedback control of dynamic systems* (Addison-Wesley, 1994).
9. Ziegler, J. G. and Nichols, N. B. *Trans. ASME* **64**, 759-768 (1942).
10. Dovichi, N. J. Personal Communication (1997).

CHAPTER 4: DIRECT ANALYSIS OF CD-PCR PRODUCTS BY CYCLE SEQUENCING

4.1 Introduction

Since the development of bisulfite-based cytosine methylation analysis [1] there have been several reports of its use[2-15]. In most of these reports the PCR products from deaminated DNA (the CD-PCR products) were cloned into a sequencing vector and several clones were sequenced to determine the extent of methylation for each cytosine residue in the sequence of interest. Cloning the CD-PCR products in order to determine their sequence has limitations. The steps involved in cloning are time-consuming and expensive. In addition, cloning PCR products can be a difficult task, often with low success rates.

The most severe limitation of cloning the CD-PCR products arises from the nature of the cloning process itself. Cloning is a sampling process: each clone arises from a single CD-PCR product molecule. Therefore, many clones need to be sequenced to obtain a statistically significant sample size to estimate methylation levels. This limitation is only problematic when the CD-PCR product contains more than one sequence. The CD-PCR product will have more than one sequence when the original genomic DNA sample is partially methylated. The concept of partially methylated DNA is described in more detail in Sections 1.3.2.1 and 1.2.2.5 and Figure 1.15.

This situation occurs when the original genomic DNA sample is partially methylated, when a specific cytosine residue is methylated in some DNA molecules and unmethylated in others. Biologically, partially methylated DNA occurs for imprinted genes[16], in tumor sample DNA[17], X-inactivation[18], and in DNA samples from across different tissues of an organism.

Due to the heterogeneous nature of DNA methylation it is preferable to avoid the cloning process and sequence the CD-PCR products directly. There are only a few reports using direct sequencing of CD-PCR products[7, 9, 10]. In principle, peak areas in sequencing electropherograms from direct sequencing of CD-PCR products can be used to estimate the extent of methylation at each cytosine in any DNA sequence. Paul and Clark have developed a method where products from sequencing reactions of CD-PCR products are analyzed on a ABI automated DNA sequencing instrument[10]. They compared peak heights in C and T lanes of a sequencing electropherogram, using this ratio to estimate the extent of methylation at cytosine residues without cloning the CD-PCR products.

In this chapter I present an improved method to estimate the extent of methylation

at any cytosine in a DNA sequence by directly cycle-sequencing the CD-PCR products. I performed single termination sequencing reactions directly from the CD-PCR products with Thermosequenase™ and analyzed the sequencing products on a capillary electrophoresis instrument. Analysis of both the precision and accuracy of this method showed that it was more precise than the method described by Paul and Clark, and required one-half the amount of sequencing lanes and reagents.

4.2 Experimental

4.2.1 Preparation of methylated DNA standard

30 µg of pUC19 plasmid DNA was linearized with EcoR1 restriction enzyme (Gibco), precipitated in 95% ethanol and resuspended in TE buffer (10 mM tris, 1 mM EDTA). 17 µg of the linear plasmid was methylated with *SssI* methylase (New England Biolabs). The 30 µL methylation reaction contained 6 units of methylase, 2 µL of restriction buffer 2 (New England Biolabs) and 2 µL of 1.6 mM s-adenosylmethionine. After 90 minutes incubation at 37°C another 2 µL of 1.6 mM s-adenosylmethionine was added to the reaction which was incubated for another 90 minutes. The DNA was checked for complete methylation by digestion with *HpaII* restriction enzyme. 1 µg of both the methylated and unmethylated DNA was digested with 10 units of *HpaII* for 1 hour and run on a 1% agarose gel.

4.2.2 Sodium bisulfite deamination of methylated and unmethylated pUC 19

The concentrations of both the methylated and the unmethylated pUC 19 were measured by fluorescence in a 5 µg/mL ethidum bromide solution with a Turner TD-700 filter fluorometer. Both DNA solutions were diluted to a concentration of 300 ng/µL. 10 µL of both DNA solutions were added to separate 1.5 mL microfuge tubes. 5 µL of 1 M NaOH was added to each tube and the tubes were incubated at 55°C for 10 minutes to ensure complete denaturation of the DNA. 1.2 mL of a freshly prepared 4 M sodium bisulfite (Sigma) and 1mM hydroquinone (Sigma) solution at pH 5.0 was added to each tube while they were at 55°C. The mixtures were covered with light mineral oil (Sigma) to prevent evaporation and incubated at 50°C. After 24 hours the tubes were cooled on ice and the oil removed. 10 µL of glassmilk (Bio101) was added to each tube and the DNA allowed to bind to the glassmilk for 20 minutes on ice. The glass beads were washed three times with NEW wash (Bio101) and the deaminated DNA was recovered in 50 µL of TE buffer. 20 µL of 1 M NaOH was added to both tubes to complete the

deamination reaction. Both tubes were incubated for 15 minutes at room temperature before 45 µL of 7.5 M ammonium acetate (pH 7) was added. The DNA was precipitated with 400 µL of 95% ethanol and resuspended in 50 µL TE.

4.2.3 Polymerase chain reaction

A 250 base pair sequence was amplified from the deamination products of both the methylated and unmethylated DNA by PCR. The PCR reactions were of 100 µL total volume, containing 100 µM of each dNTP, 1 X PCR buffer (Gibco), 1.5 mM magnesium chloride, approximately 50 pmol of each P100 and P307M13 primers, and 2.5 units of Taq polymerase (Gibco). The sequence of primers was as follows:

P100 = 5' TATGGTGTATTTTTAGTATAATTTG 3'

P307M13=5' TGTAAAACGACGGCCAGTATCATAACATTAACCTATAAAAATA 3'

These primers amplify a 250 base pair product from the deaminated pUC19 as well as incorporate a priming site for the M13-21 forward-sequencing primer at one end of the PCR product. Thermal cycling was done with a MJ Research PTC-100 with a heated lid. Cycling conditions were: 30 cycles of 94°C for 10 seconds, 58°C for 30 seconds, and 72°C for 60 seconds. 5 µL of each CD-PCR reaction product was run on a 1% agarose gel to estimate its purity.

4.2.4 Cycle sequencing of CD-PCR product standards

Both CD-PCR products were purified by a Qiagen PCR cleanup spin column (Qiagen) and their concentrations were measured by ethidium bromide fluorescence. To obtain standards for a calibration curve the CD-PCR product from the methylated and deaminated pUC19 was mixed with the CD-PCR product from the unmethylated and deaminated pUC19 in 25%, 50% and 75% ratios. Cycle sequencing was performed on the CD-PCR products from both methylated and unmethylated pUC19 as well as the CD-PCR product mixtures using the Thermosequenase kit with 7-deaza-G (Amersham). The 8 µL sequencing reactions contained 2 µL of Thermosequenase A reagent, 0.4 pmol of ROX M13-21 dye-labeled primer (Perkin Elmer), and 10 ng of CD-PCR product or CD-PCR product mixture as template. Cycling conditions were: 30 cycles of 95°C for 30 seconds and 55°C for 30 seconds. Three sequencing reactions were performed on each PCR product or PCR product mixture. After cycling, the sequencing reactions were precipitated with ethanol and resuspended in 5 µL of deionized formamide which were then ready for injection on the capillary electrophoresis instrument.

4.2.5 Separation of the sequencing fragments by CGE

The 5-capillary electrophoresis instrument with sheath flow cuvette was constructed in-house and has been described elsewhere[19]. The 543 nm line of the helium-neon laser (Melles Griot) was used to excite the ROX-labeled dye primer and fluorescence was collected through a 605DF10 bandpass filter (Omega Optical). The intensity of the fluorescence was sampled at 4 Hz. Capillaries (42 cm length x 150 μm O.D. x 50 μm I.D.) were coated with γ-methacryloxypropyltrimethoxysilane (Sigma) in 95% ethanol. The coated capillaries were filled with a polymerizing 7% acrylamide solution without crosslinker and the solution was allowed to polymerize overnight. 7 Molar urea (ICN) was present as a denaturant and 1X TBE (89mM tris, 89mM borate, 1 mM EDTA) was used as a buffer.

The capillaries were held at 40°C during the injection and separation procedures. Samples were injected on the polyacrylamide-filled capillaries by applying a 100V/cm electric field for 15 seconds. After injection, the sample vial was replaced with a 1X TBE running buffer and a 150V/cm field was applied across the capillary to separate the DNA fragments. Three sequencing reactions from each of the two pure CD-PCR's and the three CD-PCR product mixtures, 15 samples in total, were injected.

4.2.6 Data Analysis

Electropherograms from the sequencing runs were imported into Peak Fit v4.0 (Jandel Scientific) and the adenine peaks, which were complementary to the positions of two partially methylated cytosine residues, were identified. The nearest adenine peak, which was complementary to a thymine residue, was also identified. After baseline subtraction, all peaks in this area were fit with a set of 5 parameter GEMG (half Gaussian-exponentially modified Gaussian hybrid) curves to calculate their areas. The area of the adenine peak complementary to the 5-Me-C position was divided by the area of the closest neighboring adenine peak

$$Ratio = \frac{area_{5\text{-}Me\text{-}C}}{area_T} \qquad (4.1)$$

where Ratio was the ratio of the areas of the two peaks, $area_{5\text{-}Me\text{-}C}$ was the area of the adenine peak that was complementary to the 5-Me-C in the original (non-deaminated)

sequence, and $area_T$ was the area of the adenine peak that was complimentary to a neighboring thymine peak in the original sequence. The latter adenine peak acts as an internal standard for the 5-Me-C peaks. This calculation was applied to each sequencing electropherogram. This ratio was defined as 100% for the completely unmethylated sample and as 0% for the completely methylated sample. The fraction of methylation at these two sites in all of the mixed samples was estimated by

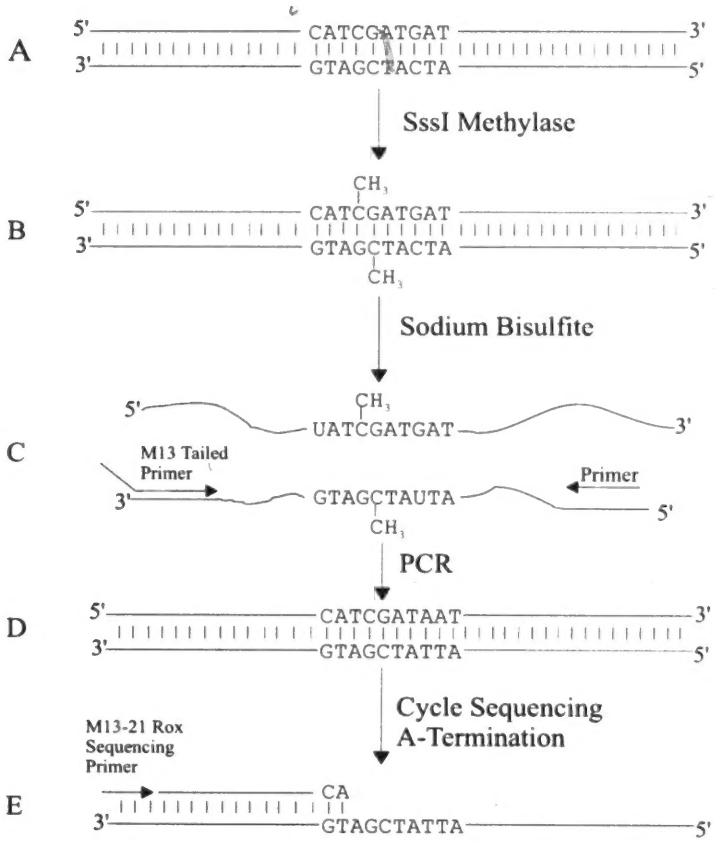
$$molefraction\ C_{sample} = \frac{Ratio_{sample} - Ratio_{0\%}}{Ratio_{100\%} - Ratio_{0\%}} = \frac{mol\ C_{final}}{mol\ C_{final} + mol\ 5\text{-}Me\text{-}C_{final}} \qquad (4.2)$$

where $Ratio_{100\%}$, $Ratio_{0\%}$, and $Ratio_{sample}$ were the peak ratios for the 100% unmethylated standard, the 0% unmethylated standard (fully methylated) and the sample, respectively. The rightmost term is the definition of mole fraction. Mol C_{final} and mol 5-Me-C_{final} are the number of moles in the CD-PCR products from the unmethylated and methylated pUC 19, respectively.

4.3 Results and Discussion

A flowchart of the chemistry for the methylated pUC 19 is shown in Figure 4.1. Double-stranded pUC 19 was first treated with SssI methylase, which quantitatively methylated all cytosines that were located 5' to a guanine. This methylated DNA was next treated with bisulfite, which quantitatively converted unmethylated cytosines to uracil, while leaving methylated cytosine unaffected. After the bisulfite treatment the two strands were no longer complementary. One strand was chosen for PCR amplification by the appropriate choice of primers. During amplification, U was converted to T and 5-Me-C was converted to C. The 5' end of one of the PCR primers incorporated a priming site for the M13-21 forward standard sequencing primer into the bottom strand of the CD-PCR product (part D of Figure 4.1). The bottom strand was then used as a template for a single base A-termination sequencing reaction. A new peak appeared in the electropherogram when unmethylated C's were converted to T's in the sequencing template, generating an A in the sequencing ladder. Positions of 5-Me-C, having been converted to C in the CD-PCR product, did not generate a peak in the A-termination sequencing reaction. Because of this, sequencing electropherograms generated from mixtures of the CD-PCR products have smaller peaks at the positions corresponding to 5-Me-C than at the positions

Figure 4.1 CD-PCR and direct cycle sequencing chemistry for methylation analysis



corresponding to T's or fully unmethylated C's.

4.3.1 Creation of methylated standard

pUC 19 DNA is a circular DNA molecule of approximately 2.6 kilobases in length. I chose pUC 19 as a standard because it is readily available and its sequence is known. SssI methylase (also known as CpG methylase), which converts all C's in a CpG dinucleotide to a 5-Me-C, was used to create the methylated DNA standard. The completeness of the methylation reaction was checked by digesting the methylated and unmethylated DNA with *HpaII* restriction enzyme, which cleaves at all 5' CCGG 3' sequences unless the second cytosine is methylated. Figure 4.2 shows the agarose gel of unmethylated pUC 19 DNA (lane 1) and SssI methylated pUC 19 DNA (lane 2) after digestion with *HpaII*. Lane 3 is a 1 kb ladder size standard (Gibco). There was no visible digestion of the methylated pUC 19 while the unmethylated pUC 19 was completely digested. Of course this method does not check for complete methylation at cytosines outside *HpaII* sites or allow the detection of very small amounts of incompletely methylated DNA. Nevertheless it did indicate that the methylase was indeed active before proceeding with the deamination chemistry.

4.3.2 Accuracy and precision of single-termination Thermosequenase chemistry

One electropherogram from the sequencing of each CD-PCR product mixture is shown in Figure 4.3. Arrows indicate peaks that corresponded to the positions of 5-Me-C in the original pUC 19 sequence. These peaks became progressively smaller as the amount of PCR product from the methylated pUC19 was increased. Areas of peaks that correspond to T's in the pUC19 sequence did not change with respect to one another as the ratio of PCR products from methylated and unmethylated pUC19 was varied. This fact allowed these peaks to serve as an excellent internal standard for determining the extent of methylation at each cytosine. In addition, small peaks, known as ghost peaks, existed in the baselines of all the electropherograms.

Equations 4.1 and 4.2 and the measured peak areas were used to calculate the observed mole fraction of CD-PCR product from the methylated pUC 19, compensating for slight variations in peak heights due to Thermosequenase chemistry as well as for the ghost peaks. Percentages of methylation obtained for peaks labeled X and Y in Figure 4.3 are reported for each standard mixture in Table 4.1. The standard deviations for the fully methylated and unmethylated samples are simply the standard deviations of the peak

Figure 4.2 Photograph of agarose gel of *HpaII* digest of unmethylated pUC 19 (lane 1) and pUC 19 methylated by CpG methylase (lane 2). Lane 3 is 1kb ladder size standard DNA.

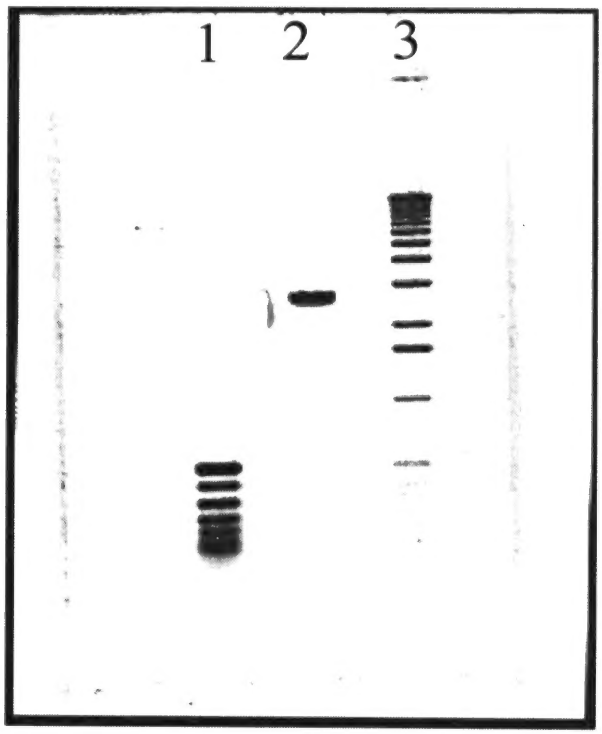


Figure 4.3 Electropherograms from A-termination sequencing of standard CD-PCR product mixtures. Percentages to the left of each electropherogram refer to the percentage of CD-PCR product from methylated DNA in each mixture. Arrows refer to positions corresponding to cytosines methylated by SssI methylase

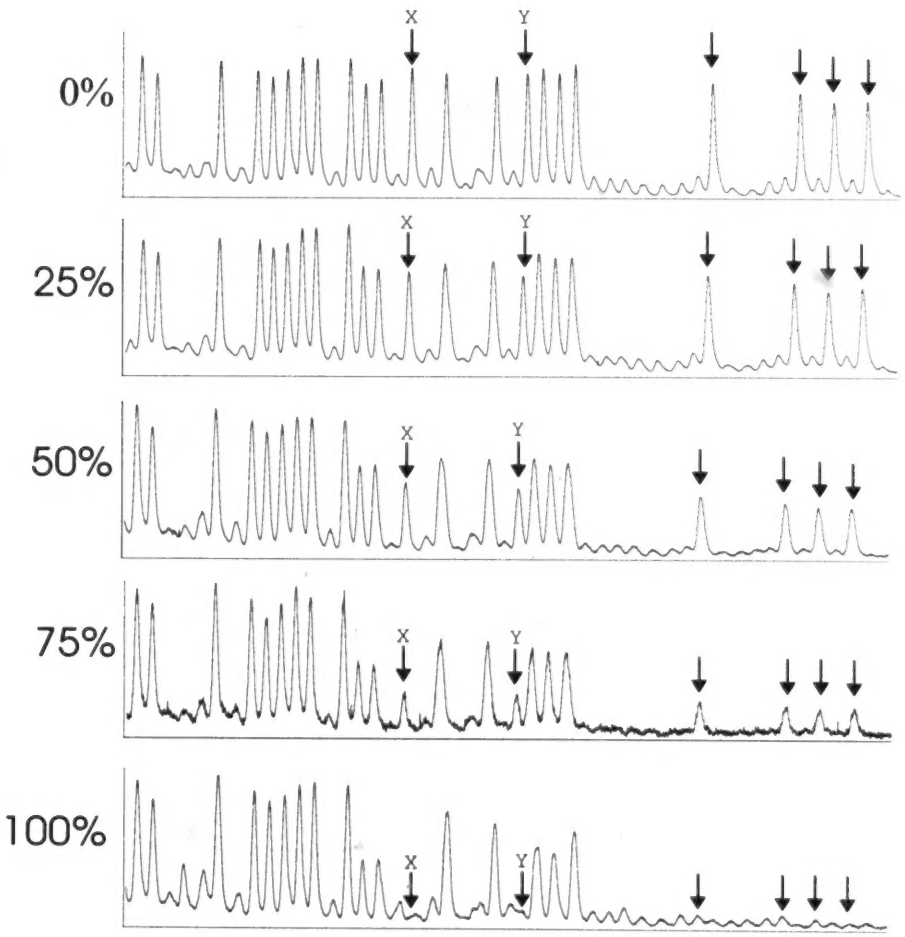


Table 4.1 Percentages of unmethylated DNA as determined by direct sequencing of standard mixtures of CD-PCR products from deaminated and methylated pUC19 with CD-PCR products from deaminated and unmethylated pUC19. Columns X and Y refer to data collected from the peaks labeled X and Y in Figure 4.4.

% methylated	X (avg)	σ_X	Y (avg)	σ_Y
0%	0 %	1.0 %	0 %	3.4 %
25%	30.1 %	3.2 %	25.2 %	7.8 %
50%	55.1 %	2.0 %	48.2 %	7.2 %
75%	78.0 %	2.9 %	76.5 %	6.8 %
100%	100 %	1.2 %	100 %	5.2 %

areas obtained from the triplicate measurements. Standard deviations for the 75%, 50%, and 25% mixtures are the overall standard deviations. The average absolute precision obtained for these measurements was 5%. The measured values for peak Y are within experimental error of the expected value, while the data for peak X tend to be slightly higher than the expected value. It is possible that the calculation of peak areas may be biased by the peak fitting methods used. It is also possible that equations 4.1 and 4.2 cannot completely compensate for subtle variations in peak areas due to Thermosequenase chemistry.

Single-termination sequencing chemistry is well suited to the direct analysis of methylcytosine using the bisulfite deamination chemistry. To perform the CD-PCR reactions the primary DNA sequence must already be known. Therefore, there is no need to perform termination reactions for all four bases. All the methylation information is contained in either the A- or T- termination reactions (depending on which strand of the PCR reaction is chosen to be sequenced). In their previous report, Paul and Clark[10] calculate the ratio of methylation at a particular cytosine by calculating the ratio of the peak height in the C lane of the sequencing gel to the peak height in the T lane. Although they do not report replicate measurements, the authors claim an error margin of 10%, which is quite good considering the variation in peak heights present due solely to Sequenase reaction chemistry. The approach of Paul and Clark provides poorer precision than the approach described in this chapter. They also require twice as many lanes of the sequencing gel, as well as twice the amount of sequencing reagents. Sequencing of short PCR products can be difficult with the conventional Sequenase system because short PCR products tend to reanneal before the Sequenase is able to replicate the entire molecule. Paul and Clark use a biotinylated primer in order to bind their PCR products to streptavidin beads to prepare a single-stranded DNA template that can be easily sequenced by Sequenase.

In contrast, short PCR products can be easily sequenced by cycle-sequencing with a thermostable polymerase (such as Taq) because the replication takes place at 70 °C, soon after the sequencing primers have annealed. Under these conditions the sequencing reactions are complete before any significant amount of template can reanneal. Unfortunately, although Taq polymerase can be used for DNA sequencing, it is not well suited to quantitative work such as methylation analysis. Taq polymerase discriminates

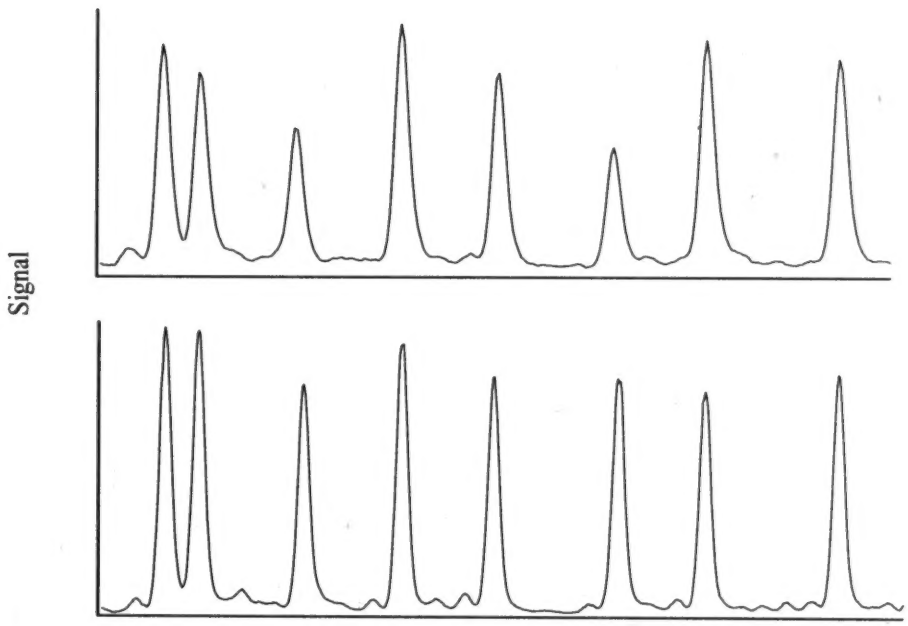
against the dideoxy nucleotide triphosphate terminators used in sequencing[20]. The result is a sequencing electropherogram with uneven peaks that would not be suitable for methylation analysis. Tabor and Richardson identified a single amino acid in Taq polymerase that is responsible for the low incorporation rate of dideoxy nucleotides[20]. A genetically engineered Taq polymerase that does not discriminate against the dideoxy terminators has been produced. It is sold as Thermosequenase™ and has proven to be a considerable advance in DNA sequencing technology. Figure 4.4 shows small sections of single termination sequencing electropherograms created with Taq polymerase (top) and Thermosequenase™ (bottom). The peak heights for in the electropherogram of sequencing fragments created by Taq polymerase have a relative standard deviation of 28%. In contrast, the peak heights from the Thermosequenase™ reaction have a relative standard deviation of 8%.

In addition to discrimination against dideoxy nucleotides, DNA polymerases also discriminate against the dye-labeled terminators used for DNA sequencing when standard dye-labeled primers are not available. Because the use of dye-terminators greatly increases the variations of peak heights in a sequencing electropherogram I chose to use dye-primer sequencing chemistry. Because dye-primers are only available for standard sequences, I added the sequence of the M13-21 forward sequencing primer to the 5' end of the PCR primers. This incorporated a binding site for the M13-21 sequencing primer at the 3' end of one strand of the CD-PCR products which allowed me to sequence the CD-PCR products directly with dye-primer Thermosequenase™ sequencing chemistry.

4.3.3 Direct CD-PCR sequencing *vs.* cloning

Direct sequencing of PCR products from deaminated DNA is considerably more cost effective in terms of both time and money than cloning the PCR products prior to sequencing. Cloning of the PCR products requires several days of time in addition to the deamination, PCR and sequencing. Cloning involves selecting single PCR product molecules, and amplifying them in a cloning vector. Because single discrete molecules are being selected the distribution of molecules selected is governed by the binomial distribution. Therefore a large number of clones must be sequenced to obtain a precise estimate of methylation levels. The relative precision obtained for a measurement made by selecting clones follows the expression:

Figure 4.4 A section of an electropherogram of T-termination sequencing fragments created with Taq polymerase (top) and Thermosequenase (bottom)



$$\sigma_{relative} = \sqrt{\frac{1-p}{N \times p}} \qquad (4.3)$$

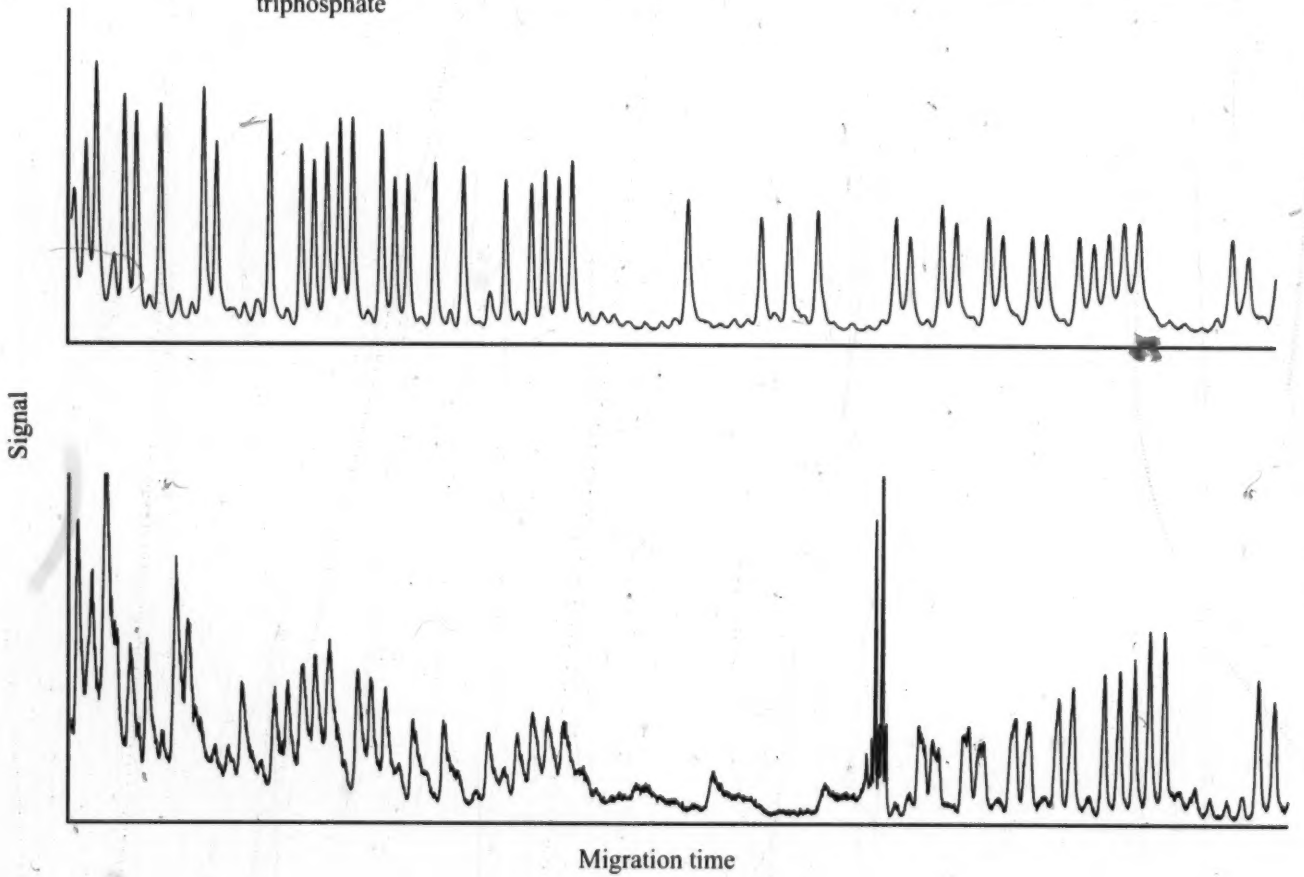
where p is the probability that a particular cytosine is methylated and N is the number of clones that have been sequenced. To obtain the 5% relative standard deviation achieved by direct sequencing for a cytosine in the 50% methylated sample, 400 clones would need to be sequenced.

Cloning the PCR products prior to sequencing has one advantage. Because cloning selects single DNA molecules it is possible to determine if methylation at two different cytosine residues are related. This situation is known to exist for some genes on the X chromosome of female mammals where it is known as X inactivation[18]. In these cases one copy of a gene is heavily methylated at cytosines on one of the X chromosomes but remains unmethylated on the other. This situation would be immediately obvious when sequencing clones; direct sequencing would indicate only that some cytosines are methylated at the 50% level.

4.3.4 Problems encountered with direct sequencing of CD-PCR products

These experiments were not without problems. Initially I was unable to obtain useable sequencing electropherograms due to a severe disruption in the separation near the beginning of the run. Fortunately, I was able to eliminate this disruption by using the Thermosequenase™ kit containing 7-deaza-deoxyguanosine triphosphate (7-deaza-G) instead of standard deoxyguanosine triphosphate. 7-deaza-G forms weaker basepairs with cytosine than standard guanosine and cannot participate in base-pairing involving hydrogen bonds through the nitrogen in the 7 position, such as Hoogsteen basepairs. It is used in DNA sequencing to reduce secondary structures in DNA sequencing fragments that result in compressions. Figure 4.5 shows electropherograms of the early part of a sequencing run of the CD-PCR product from unmethylated DNA generated with the standard Thermosequenase™ kit and the kit containing 7-deaza-G. The top trace, of sequencing fragments containing 7-deaza-G, is clean while the bottom electropherogram, of sequencing fragments with standard G, has a large depression of the peaks beginning at

Figure 4.5 Sequencing electropherograms of CD-PCR product from unmethylated DNA using Thermosequenase with 7-deaza-G (top) and Thermosequenase with standard deoxyguanosine triphosphate



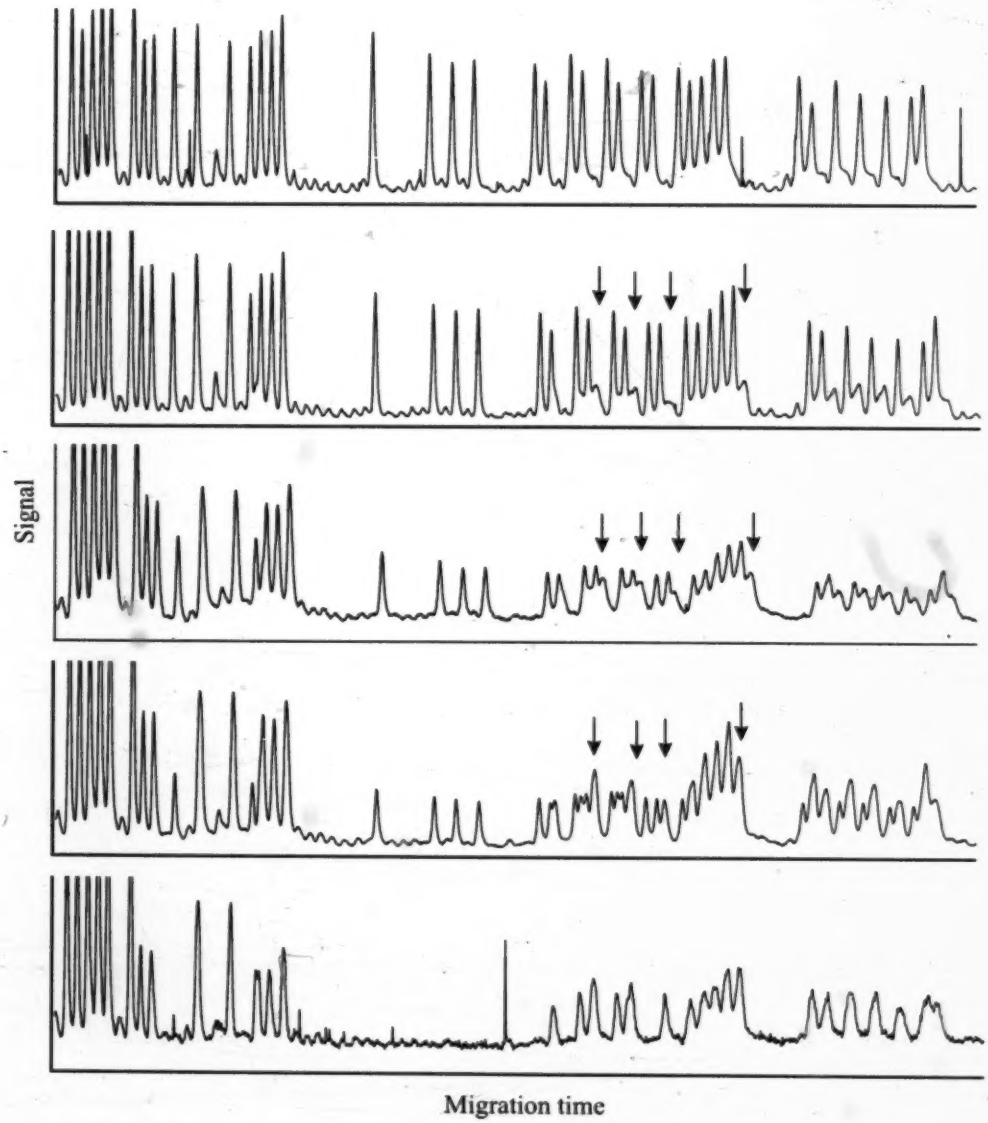
about 35 bases into the run and fully recovering at about 50 bases. I am not sure what caused this artifact but the fact that 7-deaza-G resolved it suggests that it is due to secondary structures in the sequencing fragments.

Sequencing electropherograms of the CD-PCR product mixtures displayed a more serious artifact that we were unable to resolve. The electropherograms shown in Figure 4.3 are shown again in Figure 4.6, this time with data shown to 105 bases. As the percentage of CD-PCR products from the methylated pUC 19 was increased, the resolution of the peaks worsened. Furthermore, extra peaks appeared in the later part of the electropherograms of the mixed CD-PCR products. These peaks increased in size as the percentage methylation was increased from 25% to 75%. It is possible that these extra peaks are the sequencing fragments from the methylated DNA CD-PCR product. These sequencing fragments contained G's in their sequence and might have had different mobilities in the polyacrylamide solution than sequencing fragments from the unmethylated DNA CD-PCR product, which contained no G's at all. The sudden onset of the extra peaks suggested that these mobility differences are not fundamental to this system but rather a peculiarity of this particular DNA sequence. To determine if this is true would require site-specific mutagenesis studies, is beyond the scope of this thesis. Sequence-based HLA typing experiments[21], which also generate sequencing fragments with a degenerate sequence, which are cleanly separated on a polyacrylamide slab gel. Although HLA typing experiments are less demanding than methylation analysis, they do suggest that the mobility shift problem in methylation analysis is not insurmountable.

4.4 Conclusion

Single termination dye-primer Thermosequenase™ sequencing of CD-PCR products was a precise and accurate method to quantitatively determine the methylation state of all cytosines in a DNA sequence. The precision of measurements that this method is able to achieve is equivalent to sequencing hundreds of clones while avoiding the considerable time and expense of the cloning procedures. Experiments in this chapter do not validate the CD-PCR chemistry itself; they only determine the accuracy and precision with which the populations of sequences in a CD-PCR product can be measured by direct sequencing. Validation of the CD-PCR chemistry is discussed in Chapter 5.

Figure 4.6 Electropherograms of sequencing reactions of mixtures of CD-PCR products from unmethylated pUC 19 (top) to methylated pUC 19 (bottom). These elctropherograms cover fragment sizes 20-105 bases.



4.5 Bibliography

1. Frommer, M.; McDonald, L. E.; Millar, D. S.; Collis, C. M.; Watt, F.; Grigg, G. W.; Molloy, P. L. and Paul, C. L. *Proceedings of the National Academy of Sciences, USA* **89**, 1827-1831 (1992).
2. Clark, S. J.; Harrison, J.; Paul, C. and Frommer, M. *Nucleic Acids Research* **22**, 2990-2997 (1994).
3. Clark, S. J.; Harrison, J. and Frommer, M. *Nature Genetics* **10**, 20-27 (1995).
4. Feil, R.; Charlton, J.; Bird, A. P.; Walter, J. and Reik, W. *Nucleic Acids Research* **22**, 695-696 (1994).
5. Grigg, G. and Clark, S. *BioEssays* **16**, 431-436 (1994).
6. Meyer, P.; Niedenhof, I. and Lohuis, M. *The EMBO Journal* **13**, 2084-2088 (1994).
7. Myohanen, S.; Wahlfors, J. and Janne, J. *DNA Sequence* **5**, 1-8 (1994).
8. Olek, A.; Oswald, J. and Walter, J. *Nucleic Acids Research* **24**, 5064-5066 (1996).
9. Park, J.-G. and Chapman, V. M. *Molecular and Cellular Biology* **14**, 7975-7983 (1994).
10. Paul, C. L. and Clark, S. J. *BioTechniques* **21**, 126-133 (1996).
11. Raizis, A. M.; Schmitt, F. and Jost, J. P. *Analytical Biochemistry* **226**, 161-166 (1995).
12. Reeben, M. and Prydz, H. *BioTechniques* **16**, 416-417 (1994).
13. Rother, K. I.; Silke, J.; Georgiev, O.; Schaffner, W. and Matsuo, K. *Analytical Biochemistry* **231**, 263-265 (1995).
14. Sadri, R. and Hornsby, P. J. *Nucleic Acids Research* **24**, 5058-5059 (1996).
15. Selker, E. U.; Fritz, D. Y. and Singer, M. J. *Science* **262**, 1724-1728 (1993).
16. Monk, M. *Developmental Genetics* **17**, 188-197 (1995).
17. Laird, P. W. and Jaenisch, R. *Annual Review of Genetics* **30**, 441-464 (1996).
18. Moore, T.; Hurst, L. D. and Reik, W. *Developmental Genetics* **17**, 206-211 (1995).
19. Zhang, J. in *The Sheath Flow Cuvette in High-Sensitivity Fluorescence Detection for DNA Sequencing in Single and Multiple Capillary Systems* Ph.D thesis, Department of Chemistry, University of Alberta, Edmonton (1994).
20. Tabor, S. and Richardson, C. C. *Proceedings of the National Academy of*

Sciences, USA **92**, 6339-6343 (1995).

21. Rozemuller, E. H.; Eliaou, J. F.; Baxter-Lowe, L. A.; Charron, D.; Kronick, M. and Tilanus, M. G. J. *Tissue Antigens* **46**, 96-103 (1995).

CHAPTER 5: COMBATING PCR BIAS IN BISULFITE-BASED CYTOSINE METHYLATION ANALYSIS: BETAINE-MODIFIED CD-PCR

5.1 Introduction

As with methylation sensitive restriction endonucleases[1], bisulfite-based cytosine deamination-PCR (CD-PCR) chemistry can provide quantitative information about the methylation state of cytosines[2]. In addition to providing methylation information for cytosines that the restriction endonuclease-Southern blotting methodology cannot analyze, CD-PCR chemistry is much more sensitive than Southern blotting. CD-PCR chemistry has been used to determine the methylation status of genomic DNA samples from as few as one hundred cells[3]. Even better sensitivity than this will be required to determine methylation levels of embryos and zygotes. Such experiments will be necessary to elucidate the role of methylation in growth and development.

The key to the extraordinary sensitivity of the CD-PCR chemistry is the polymerase chain reaction (PCR)[4]. The PCR works by exponentially amplifying any short DNA sequence between a set of oligonucleotide primers. Unfortunately, PCR chemistry is not quantitative. It is difficult to accurately determine the original concentration of a DNA sequence based on the concentration of its PCR product.

Nevertheless, because there is no technology that rivals the sensitivity of the PCR, much effort has been expended to make PCR amenable to quantitation. Most of these methods involve competitive PCR, where a known amount of a DNA is added to the samples as an internal standard. The internal standard DNA sequence has the same primer binding sites as the analyte DNA sequence and so competes with the analyte DNA sequence in the extent of amplification. The amount of internal standard added is varied until the concentration of its PCR product is equal to the concentration of the analyte PCR product. The internal standard DNA sequence must be chosen to match the analyte DNA sequence as closely as possible[5], often differing from the analyte sequence by only a restriction site or a small deletion somewhere in the middle of the sequence. However, even with a well chosen internal standard, competitive PCR experiments often only provide accuracy within an order of magnitude. There is a vast amount of literature dealing with quantitative PCR for measuring mRNA levels and retroviral loads which is beyond the scope of this thesis. However, a good review can be found in a PCR methodology textbook[6].

Because CD-PCR chemistry relies on the PCR reaction to extract the methylation information from deaminated DNA, it is not necessarily a quantitative analytical technique. Fortunately, in a methylation analysis experiment it is not necessary to measure the absolute amount of methylated DNA in a sample, only the fraction of the DNA that is methylated. In addition, because primers can be designed to amplify sequences from both methylated and unmethylated DNA, a CD-PCR experiment is a competitive PCR experiment; the DNA sequence from the methylated and deaminated DNA competes for primers with the DNA sequence from the unmethylated and deaminated DNA.

Unfortunately, the inherently competitive nature of CD-PCR does not allow the experimenter to choose the nature of the competing DNA sequences. That is decided by the primary DNA sequence and the methylation state of the DNA, the latter of which is unknown when designing the experiment. The methylated and unmethylated DNA CD-PCR products, having different primary DNA sequences, can also have different rates of amplification in a PCR reaction. This means that CD-PCR chemistry, while being potentially quantitative, is not necessarily so.

Surprisingly, while quantitative estimates of methylation levels obtained by CD-PCR chemistry have been reported[7-9], these reports did not present a calibration curve or otherwise validate the accuracy of the CD-PCR chemistry. These studies even report percentages of methylation to three significant figures. Data in these reports was obtained by sequencing clones, in which the sampling process introduces significant uncertainty. In addition, the authors make a more serious error in assuming that CD-PCR chemistry has a linear response to methylation levels. The importance of an accurate methylation assay must not be underestimated; methylation levels change during development and tumorgenesis[10, 11]. Because the technology to do methylation assays on single cells does not yet exist, one must measure the partial methylation levels in samples of cells undergoing differentiation or carcinogenesis. Non-quantitative assays are useless in this regard.

In this chapter I show that CD-PCR does not necessarily have a linear response to methylation levels. DNA sequences arising from methylated and deaminated DNA are

discriminated against in the PCR, leading to measured methylation levels that underestimate the actual methylation levels. Furthermore, I demonstrate that this PCR bias leads to maximum errors when the initial DNA sample is 50% methylated, a situation that occurs naturally for imprinted genes[11, 12]. However, the addition of betaine (N,N,N-trimethylglycine) to the PCR step in the CD-PCR chemistry, which I call betaine-modified CD-PCR, can nearly eliminate this bias.

5.2 Experimental

5.2.1 Creation of DNA standards with known methylation level

A 300 ng/μL methylated pUC 19 was made as described in Section 4.2.1. A 300 ng/μL of unmethylated pUC 19 was made by diluting a linear pUC 19 stock DNA solution. The 300 ng/μL solutions of methylated and unmethylated pUC19 DNA were mixed in 6 tubes containing 100%, 80%, 60%, 40%, 20%, and 0% of the unmethylated pUC19 respectively. Each tube contained 3 μg of total DNA. The DNA in each tube was subjected to deamination conditions identical to those described in Section 4.2.2. The final deaminated DNA solutions were also 50 μL total volume.

5.2.2 Calibration curve for CD-PCR chemistry

Each of the six deaminated standards was amplified by PCR using conditions exactly as described in Section 4.2.3. An additional set of CD-PCR products was created with identical PCR conditions except that the PCR reactions also contained 2 M betaine. Each CD-PCR product was sequenced using the Thermosequenase™ kit with 7-deaza-G as described in Section 4.2.4. The A-termination sequencing reaction products were separated by capillary gel electrophoresis as described in Section 4.2.5 and the same peaks labeled X and Y in Figure 4.4 were identified. The methylation levels for these peaks were calculated as described in Section 4.2.6.

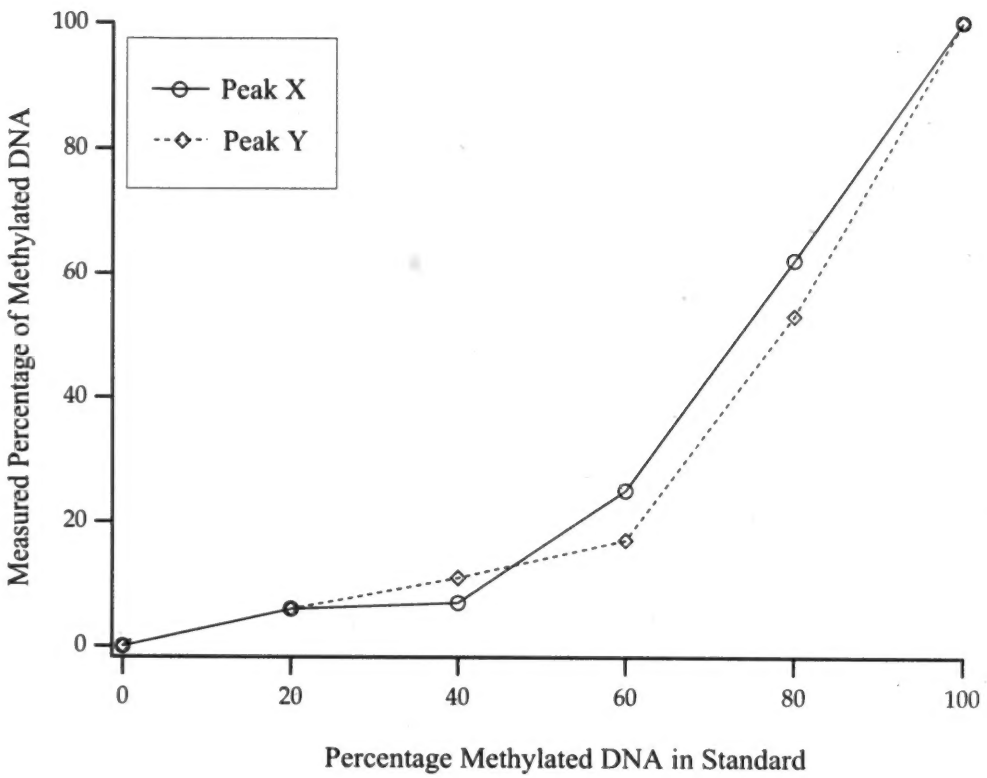
5.3 Results and discussion

Percentages of methylation in each standard as measured by conventional CD-PCR are reported in Table 5.1. A plot of measured percentage of methylation against the percentage of methylated DNA in the standards is included in Figure 5.1. Ideally, this calibration curve should be a straight line with a slope of 1. Obviously, there was bias against the methylated DNA in the CD-PCR chemistry. The error in the estimate of

Table 5.1 Calibration curve generated with conventional CD-PCR

% Methylated (standard)	Peak X	Peak Y
0	0%	0%
20	6%	6%
40	7%	11%
60	25%	17%
80	62%	53%
100	100%	100%

Figure 5.1 Measured percentage of methylalted DNA in standard by CD-PCR vs actual percentage methylated. Circles and solid line are for peak X. Diamonds and dashed line are for peak Y.



methylation was greatest at intermediate levels of methylation. For instance, in the case of an imprinted gene, where one chromosome is methylated and the other chromosome is unmethylated, an estimate of methylation levels should indicate methylation at the 50% level. If the pUC 19 sequence used in this experiment were part of that gene, I would have estimated its methylation level at approximately 18%. This error is unacceptable.

This bias was likely due to one of the following: selective destruction of the methylated DNA during the deamination reaction, systematic errors in direct cycle sequencing and peak area measurement, and/or differential amplification of the methylated and unmethylated DNA sequences during the PCR reaction. While destruction of the DNA does occur during the bisulfite deamination, it is primarily due to depurination that is accelerated by the low pH required for deamination[13]. As the methylated and unmethylated DNA sequences did not differ in the number and position of purine (A and G) residues it is unlikely that depurination was the cause of the observed bias. In Chapter 4 I showed that direct cycle-sequencing was a precise and accurate method to analyze CD-PCR products. Bias during the PCR reactions seems the most likely cause of the observed error. I decided to model the bias in the PCR. The differential amplification, D, describes the relative increase in unmethylated DNA sequence during the PCR:

$$\frac{mol\ C_{final}}{mol\ 5\text{-}Me\text{-}C_{final}} = D \times \frac{mol\ C_{initial}}{mol\ 5\text{-}Me\text{-}C_{initial}} \qquad (5.1)$$

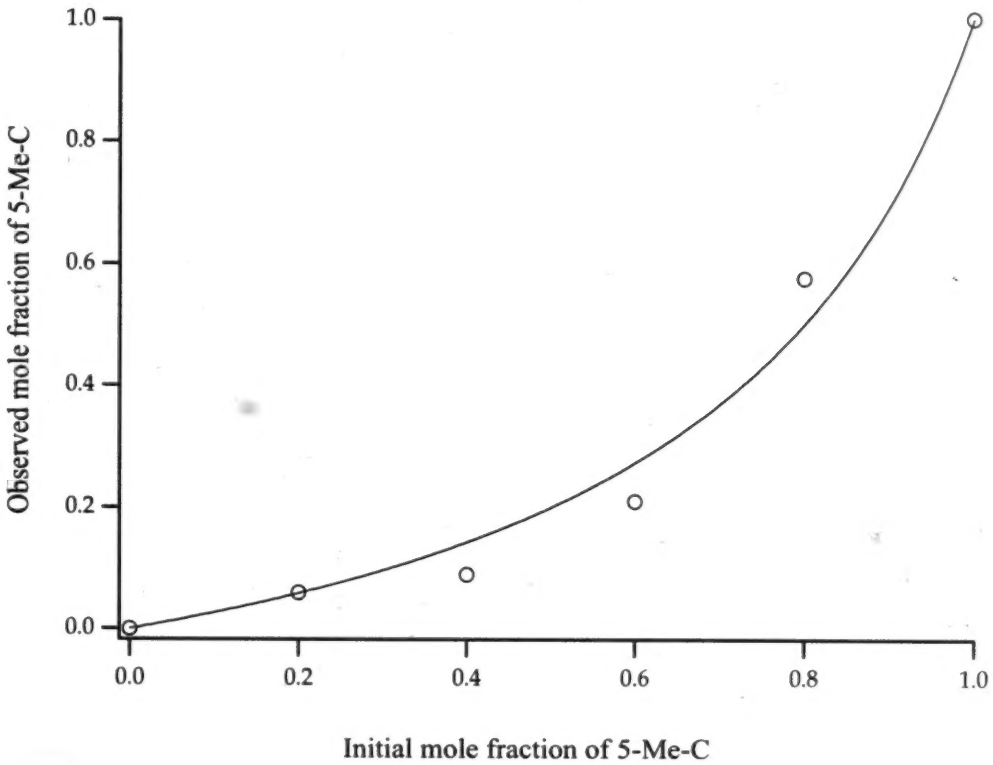
Equation 5.1 can be substituted into Equation 4.2 and rearranged to give:

$$mole\ fraction\ 5\text{-}Me\text{-}C_{observed} = \frac{mole\ fraction\ 5\text{-}Me\text{-}C_{initial}}{D + mole\ fraction\ 5\text{-}Me\text{-}C_{initial} \times (1-D)} \qquad (5.2)$$

The data for peaks X and Y were averaged and plotted in Figure 5.2; the smooth curve is the non-linear least squares fit of Equation 5.2 to the data. In this case D=4.0; that is, the unmethylated fraction of the sample was amplified four times more than the methylated fraction was during the PCR.

This result was not unexpected. The two sequences in this competitive PCR were

Figure 5.2 Plot of observed mole fraction of methylated DNA vs. initial mole fraction of methylated DNA for conventional CD-PCR. Circles are averages of data points for peaks X and Y. Smooth curve is non-linear least squares fit of equation 5.2 to the data set.



considerably different in composition. The deamination product of methylated DNA contained 5-methylcytosine residues whereas the deamination product from the unmethylated DNA contained no cytosine residues at all. The most likely cause of bias in the PCR reactions was the presence of secondary structures in the sequence of the methylated and deaminated DNA that were not present in the unmethylated and deaminated sequence. The GC base pair is more stable than the AT base pair under normal conditions. Also, cytosine is an important base for stabilizing hairpins and other types of intrastrand secondary structures[14]. Furthermore, 5-methylcytosine is even better than cytosine at stabilizing some intrastrand secondary structures[15, 16].

PCR is a kinetically driven process. The double-stranded DNA molecule to be amplified is more energetically stable than is each of its individual strands annealed to the PCR primer. PCR works because the primer anneals to the template strand before the template can reanneal to its full-length complement. DNA annealing is a bimolecular process in which the rate-limiting step is called nucleation, when short regions of complementary DNA sequences come together. Once the proper nucleation site forms, the rest of the double stranded DNA molecule reanneals rapidly. In a PCR amplification the competing kinetics are:

$$\frac{d[P_{primer}{\cdot}T_{template}]}{dt} = k_1[P_{primer}][T_{template}]t \tag{5.3}$$

$$\frac{d[T_{template}{\cdot}T_{template}]}{dt} = k_2[T_{template}]^2 t \tag{5.4}$$

where P_{primer} is the single-stranded primer, $T_{template}$ is the single-stranded template strand, $P_{primer}{\cdot}T_{template}$ is the primer annealed to the template and $T_{template}{\cdot}T_{template}$ is the template strand annealed to its complementary strand. Because they depend mostly on nucleation and not on the size of the DNA, the rate constants k_1 and k_2 in equations 5.3 and 5.4 are similar in magnitude. Only the kinetics described by equation 5.3 lead to amplification for

any particular cycle of PCR. The reaction progresses because the concentration of primer, typically micromolar, is much larger than the template concentration, which may be as small as a single molecule. As the number of cycles increases the concentration of the template increases exponentially and the primer concentration decreases exponentially until the reaction can progress no further.

The equation for the formation of intrastrand secondary structure in the template is:

$$\frac{d[T_{template\ intrastrand}]}{dt} = k_3[T_{template}]t \qquad (5.5)$$

where $T_{template\text{-}intrastrand}$ is any intrastrand secondary structure in the template DNA molecule. In this case the rate constant, k_3, can be relatively large because the sequences that hybridize to form the secondary structure are part of the same DNA strand and don't have to rely on diffusion to bring complementary strands together. If this secondary structure interferes with the primer binding site or forms a structure that cannot be relaxed by the DNA polymerase as it replicates the DNA, then that molecule will not be amplified by PCR. The extent of inhibition of amplification depends upon what type of structures can form, and how fast they form. In a non-quantitative PCR, intrastrand secondary structures are not a problem unless they inhibit the PCR to such an extent that no product is formed. This sometimes happens when PCR is used to amplify a very GC rich DNA sequence[17].

I wished to minimize the bias in CD-PCR chemistry. Assuming that secondary structures in the methylated DNA are inhibiting its amplification efficiency during the PCR, I felt that the best way to minimize the PCR bias would be to prevent the formation of the GC dependent secondary structures. Tetraalkylammonium salts, such as tetramethylammonium (TMA) and tetraethylammonium (TEA), have an isostabilizing effect on DNA[18]. These compounds are weak chaotropic agents that have a general destabilizing effect on double-stranded DNA. However, TMA and TEA also bind to AT base pairs through the major groove of double-stranded DNA[19], providing some

stabilizing effect for AT base pairs so that at high concentrations (~ 3 M) the melting temperature of the DNA is independent of base pair composition. The overall effect of TMA and TEA is preferential destabilization of GC base pairs. It is the isostabilizing effect of TMA that I desired to reduce the bias in CD-PCR chemistry. Low concentrations of TMA have been used to enhance the specificity of PCR[20]; unfortunately the addition of more than small amounts of TMA inhibits the PCR.

Betaine (N,N,N-trimethylglycine) is similar to TMA; it has isostabilizing properties at 5.2 M concentration and does not disrupt enzyme activity or DNA-protein interactions[21]. The structure of betaine is shown in Figure 5.3. Betaine is a zwitterion which does not contribute to the ionic strength in a reaction. Consequently, I felt that high concentrations of betaine might be tolerated by PCR and might help to eliminate the bias in CD-PCR.

The results of the methylation measurement of peaks X and Y using the betaine modified CD-PCR protocol are presented in Table 5.2. Data for peaks X and Y are averaged and plotted in Figure 5.4. The smooth curve in Figure 5.4 is the non-linear least squares fit of the data to equation 5.2. In this case the value of D = 1.2. The addition of betaine reduced the PCR bias by more than a factor of three. The residual bias was within the error of the direct sequencing method described in chapter 4.

Since this work was completed, a few other reports have used betaine in PCR to obtain even amplification of GC repeat regions[22] and to avoid false negatives due to preferential amplification in PCR-based HLA typing[23]. Another report claims that betaine eliminates pausing of DNA polymerases at specific sequences[24]. The authors claim that it is polymerase pausing, and not secondary structures that inhibits the PCR of GC rich sequences. Not only is betaine useful in methylation analysis but is likely to find use in other types of competitive PCR and DNA sequencing.

5.4 Conclusion

CD-PCR chemistry is a highly sensitive technique that has the potential to quantitatively determine the extent of methylation at any cytosine in a DNA sample. CD-PCR is quantitative if the methylated and unmethylated DNA sequences are amplified at equal rates. I have shown that these amplification rates are not necessarily equal; the PCR

step preferentially amplified the sequence from the unmethylated DNA. This suggests that reports of quantitation by CD-PCR which claim to measure partial methylation, but do not provide a calibration curve, are unreliable. All CD-PCR based analysis of partially methylated DNA must be checked for PCR bias. Fortunately, when PCR bias is found it can be reduced by the addition of betaine to the PCR reactions.

Figure 5.3 Structure of betaine

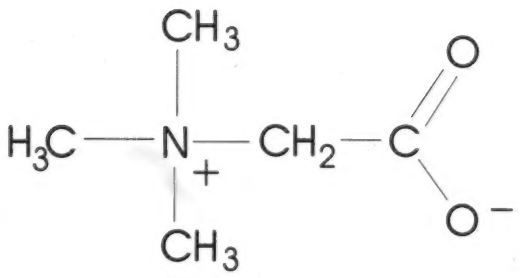
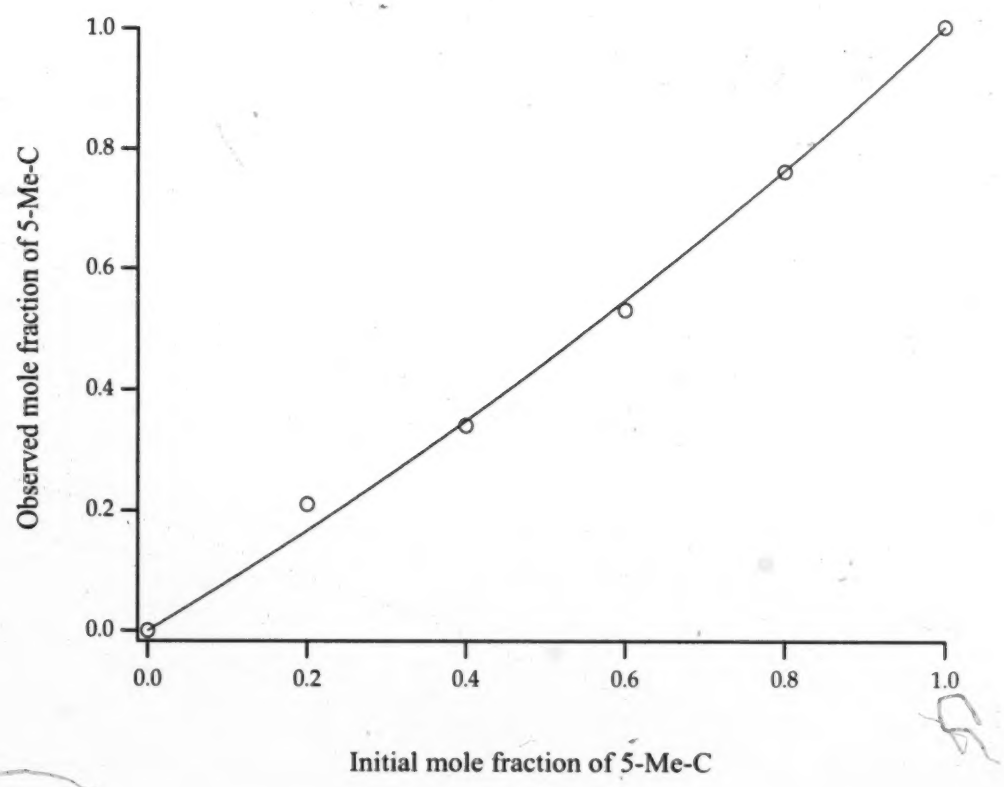


Table 5.2 Calibration curve generated with betaine modified CD-PCR

% Methylated (standard)	Peak X	Peak Y
0	0%	0%
20	18%	24%
40	40%	28%
60	58%	49%
80	79%	72%
100	100%	100%

Figure 5.4 Plot of observed mole fraction of methylated DNA vs. initial mole fraction of methylated DNA for betaine modified CD-PCR. Circles are averages of data points for peaks X and Y. Smooth curve is non-linear least squares fit of equation 5.2 to the data set.



5.5 Bibliography

1. Bird, A. P. and Southern, E. M. *Jounal of Molecular Biology.* **118**, 27-47 (1978).
2. Frommer, M.; McDonald, L. E.; Millar, D. S.; Collis, C. M.; Watt, F.; Grigg, G. W.; Molloy, P. L. and Paul, C. L. *Proceedings of the National Academy of Sciences, USA* **89**, 1827-1831 (1992).
3. Olek, A.; Oswald, J. and Walter, J. *Nucleic Acids Research* **24**, 5064-5066 (1996).
4. Mullis, K. *Cold Spring Harbor Symposia on Quantitative Biology* **51 Pt 1**, 263-273 (1986).
5. Tang, J.; Lagace, G. and Colly, R. *BioTechniques* **21**, 378-380 (1996).
6. White, B. A. *PCR Protocols: Current Methods and Applications* (Humana Press, Totowa, New Jersy, 1993).
7. Meyer, P.; Niedenhof, I. and Lohuis, M. *The EMBO Journal* **13**, 2084-2088 (1994).
8. Clark, S. J.; Harrison, J. and Frommer, M. *Nature Genetics* **10**, 20-27 (1995).
9. Park, J.-G. and Chapman, V. M. *Molecular and Cellular Biology* **14**, 7975-7983 (1994).
10. Laird, P. W. and Jaenisch, R. *Annual Review of Genetics* **30**, 441-464 (1996).
11. Monk, M. *Developmental Genetics* **17**, 188-197 (1995).
12. Moore, T.; Hurst, L. D. and Reik, W. *Developmental Genetics* **17**, 206-211 (1995).
13. Raizis, A. M.; Schmitt, F. and Jost, J.-P. *Analytical Biochemistry* **226**, 161-166 (1995).
14. Zhu L.; Chou, S. H. and Reid, B. R. *Proceedings of the National Academy of Sciences, USA* **93**, 12154-12164 (1996).
15. Murchie, A. I. H. and Lilley, D. M. J. *Nucleic Acids Research* **15**, 9641-9654 (1989).
16. Hagerman, P. J. *Biochemistry* **29**, 1980-1983 (1990).
17. Agarwal, R. K. and Perl, A. *Nucleic Acids Research* **21**, 5283-5284 (1993).
18. Melchior, W. B. and von Hippel, P. H. *Proceedings of the National Academy of Sciences, USA* **70** (1973).

19. Shapiro, J. T.; Stannard, B. S. and Felsenfeld, G. *Biochemistry* **8**, 3233-3241 (1969).

20. Hung, T.; Mak, K. and Fong, K. *Nucleic Acids Research* **18**, 4953 (1990).

21. Rees, W. A.; Yager, T. D.; Korte, J. and Hippel, P. H. v. *Biochemistry* **32**, 137-144 (1993).

22. Baskaran, N.; Kandpal, R. P.; Bhargava, A. K.; Glynn, M. W.; Bale, A. and Weissman, S. M. *Genome Research* **6**, 633-638 (1996).

23. Weissensteiner, T. and Lanchbury, J. S. *BioTechniques* **21**, 1102-1108 (1996).

24. Mytelka, D. S. and Chamberlin, M. J. *Nucleic Acids Research* **24**, 2774-2781 (1996).

CHAPTER 6: METHYLATION ANALYSIS OF THE P16/CDKN2 GENE BY CD-PCR

6.1 Introduction

Recently, considerable attention has been focused on the possibility of DNA methylation as a mechanism for the inactivation of tumor suppressor genes, and ultimately a cause of cancer. Methylation of CpG islands is correlated with transcriptional silencing of several tumor suppressor genes in tumor cells including the Rb[1,2], VHL[3], P16/CDKN2[4-6], Myf3[7], and E-cadherin[8] genes. It is believed that (cytosine 5-)-methyltransferase, with a *de novo* activity that is 1% of its maintenance activity *in vitro*[9], slowly and randomly *de novo* methylates cytosines in the CpG islands of tumor suppressor genes. Once a CpG site is methylated, the methylation is inherited by all subsequent daughter cells due to the high maintenance activity of the methyltransferase. Over time the methylation may build up to some level that would reduce or silence the transcription of that gene.

The p16/CDKN2 gene has been an active target for research in this area. The p16 protein binds to cyclin-dependent kinase 4 (CDKN4) and inhibits the binding of CDKN4 to cyclin D1, which prevents the passage through the G1 phase of the cell cycle[10]. This human p16 gene is located on chromosome 9p21, a region often deleted in cancer cells[11]. Considerable sequence analysis of the p16 gene in many primary tumors uncovered only one mutation[12]. The low rate of mutations in primary tumors suggests that a second tumor suppressor gene resides at this locus or perhaps that the gene is inactivated by means other than point mutations in most tumors.

To date, all studies of methylation of the CpG island of the p16 gene have relied on methylation sensitive restriction enzymes. Of the approximately 41 CpG dinucleotides in the CpG island of the p16 gene, only 6 sites have actually been analyzed. A recent report examined the p16 gene of 70 primary tumors[12]. In cases where only one copy of the p16 gene exists or where there is no detectable expression of the p16 transcript (21 tumors total), the authors also determined the methylation state by restriction digestion. The authors claimed that methylation is not important in silencing the p16 gene; no tumors were found to be fully methylated and only three tumors were found to be partially methylated. Partial methylation is this case means that they found methylation at one restriction site but not at other restriction sites in the same tumor sample. In this

work the authors used partial methylation in a qualitative sense as discussed in Section 1.3.2.1. The hybrid restriction enzyme/PCR system the authors used to detect methylation was not capable of quantitatively determining the methylation status of any individual cytosine. In cases where methylation of a restriction site was detected, the unmethylated restriction site was detected as well. The authors claim that this data is due to subpopulations of tumor cells with different methylation status in their sample. The authors speculate that methylation of only a few of the CpG sites is required to silence the gene.

This type of data, which the authors themselves admit is very difficult to interpret, highlights the need for a quantitative CD-PCR based methylation assay for p16. I have developed this assay and have applied it to the methylation analysis of two cell lines. I have determined that one of these cell lines was methylated at all 15 CpG positions within the 155 bases that were analyzed and the other cell line was completely unmethylated.

6.2 Experimental

All primers were designed using the sequence of the first exon of the human p16/CDKN2/MST1 gene (GeneBank accession number U12818).

6.2.1 Preparation of cell line DNA

A431 human skin carcinoma cell line cells, passage number 12, were grown to 100% confluence. HT29 human colorectal carcinoma cell line cells, passage number 18, were grown to 70% confluence. Both cultures were trypsinized from their flasks and washed twice with phosphate buffered saline (PBS). After washing both cell lines were resuspended in 300 µL of PBS. To each cell suspension was added 500 µL of digestion buffer (100 mM NaCl, 10 mM Tris-HCl, 25 mM EDTA, 0.5% sodium dodecylsulfate, 0.1 mg/mL proteinase K at pH 8.0) and the solutions were incubated for 18 hours at 50 °C on a rotor wheel. Following incubation both digests were split into two tubes. Each tube was extracted three times with 500 µL of a 50% phenol/50% chloroform solution, precipitated with 2 volumes of 95% ethanol at room temperature, and resuspended in 50 µL of TE (10 mM Tris, 1mM EDTA at pH 8.0). 1 µL of a 10 mg/mL Ribonuclease A solution was added to each tube incubated for 1 hour at 37 °C. The DNA concentration in each tube

was measured by fluorescence in a 0.5 mg/mL ethidium bromide solution. Each A431 tube contained approximately 850 ng/μL of DNA. Each HT29 tube contained approximately 300 ng/μL of DNA.

6.2.2 Confirmation of the existence of p16 in these cells lines

It was possible that a homozygous deletion had eliminated both copies of the p16 gene in the cell lines. I tested for the existence of the p16 gene by PCR. As previously described[4], primers were designed to amplify a 350 base product from exon 1 of the p16 gene if the gene was present. The PCR reactions were 50 μL total volume, containing 100 μM of each dNTP, 1 X PCR buffer (Gibco), 1.5 mM magnesium chloride, 2.5 pmol of each PE1s and PE4a primers, 1.5 μL of formamide, 1 μL of cell line DNA from section 6.2.1, and 2.5 units of Taq polymerase (Gibco). The primer sequences were as follows: PE1s = 5' GAAGAAAGAGGAGGGGCTG 3',
PE4a = 5' GCGCTACCTGATTCCAATTC 3'. Thermal cycling was done with a MJ Research PTC-100 thermal cycler with a heated lid. Cycling conditions were: 35 cycles of 94°C for 10 seconds, 60°C for 30 seconds, and 72°C for 60 seconds. 5 μL of each CD-PCR reaction product was run on a 1% agarose gel. A single band of approximately 350 bases size was seen for both the A431 and HT29 cell lines, indicating that at least one copy of the p16 gene was present in each of these cell lines.

6.2.3 Deamination of cell line DNA

5 μg of both the A431 and HT29 DNA was linearized with EcoR1 restriction enzyme, precipitated with ethanol and resuspended in 50 μL of TE. Both DNA solutions were deaminated with sodium bisulfite using conditions exactly as described in Section 4.2.2.

6.2.4 Polymerase chain reaction

A 317 base pair sequence was amplified from the deamination products of both the A431 and HT29 DNA with the PCR. The PCR reactions were 50 μL total volume, containing 100 μM of each dNTP, 1 X PCR buffer (Gibco), 1.5 mM magnesium chloride, 2.5 pmol of each c13 and c320 primers, 1.5 μL of formamide, 2 μL of deaminated cell line DNA from section 6.2.3 and 2.5 units of Taq polymerase (Gibco). The primers were are as follows: c13 = 5' AGGGGTTGGTTGGTTATTAGAGGGTGGGG 3',

C320 = 5' CCCTACAAACTT(AG)TCCTCCAAAATC 3'.

Cycling conditions were: 35 cycles of 94°C for 10 seconds, 60°C for 30 seconds, and 72°C for 60 seconds. These primers amplify a 317 base pair product from the deaminated human DNA carrying the p16 gene as the outer step in a nested PCR amplification. 10 µL of each PCR product was run on a 1% agarose gel. Only very faint bands were seen at 320 bases, no bands were seen elsewhere in the gel.

A 154 base pair sequence was amplified from the outer step PCR product by an inner step PCR. The PCR reactions were 50 µL total volume, containing 100 µM of each dNTP, 1 X PCR buffer (Gibco), 1.5 mM magnesium chloride, 2.5 pmol of each c118M13 and c272 primers, 2 µL of PCR product from the outer PCR step and 2.5 units of Taq polymerase (Gibco). The sequence of primers was as follows:

C118m13 = 5'TGTAAAACGACGGCCAGTGGGGAGTAGTATGGAGTTTT 3',

C272 = 5'CTACAAACCCTCTACCCACCTAAATC3'

Cycling conditions were: 35 cycles of 94°C for 10 seconds, 55°C for 30 seconds, and 72°C for 60 seconds. These primers amplify a 155 base pair product from the outer PCR step product above, also incorporating a priming site for the M13-21 standard sequencing primer at one end of the PCR product. 5 µL of each PCR product was run on a 1% agarose gel. Very strong bands were seen at approximately 160 base size, no bands were seen elsewhere in the gel.

6.2.5 Cycle sequencing of CD-PCR products

Both inner step CD-PCR products were purified by a Qiagen PCR cleanup spin column (Qiagen) and their concentrations measured by ethidium bromide fluorescence. Cycle sequencing was preformed on the CD-PCR products from both the A431 and HT29 cell line DNA using the Thermosequenase kit with 7-deaza-G (Amersham). The 8 µL sequencing reactions contained 2 µL of Thermosequenase T reagent, 0.4 pmol of ROX M13-21 dye-labeled primer (Perkin Elmer), and 10 ng of CD-PCR product as template. Cycling conditions were: 30 cycles of 95°C for 30 seconds and 55°C for 30 seconds. After cycling, the sequencing reactions were precipitated with ethanol and resuspended in 5 µL of deionized formamide which were then ready for injection on the capillary electrophoresis instrument.

6.2.5 Separation of the sequencing fragments

The sequencing separations were performed as described in section 4.2.5 except that they were carried out at 45°C instead of 40°C.

6.3 Results and discussion

The electropherograms from the T-terminated sequencing reactions of the CD-PCR products from both the A431 and HT29 cell lines are shown in Figure 6.1. The top panel shows migration times from 35 to 60 minutes and the bottom panel is the migration time from 60 to 85 minutes. The top (solid) trace is the data from the A431 cell line, the bottom (dotted) trace is data from the HT29 cell line. Arrows indicate positions that correspond to cytosines in CpG dinucleotides. In the A431 electropherogram there is a full height peak under each arrow indicating that the A431 cell line is fully unmethylated at all cytosines in this 150 base region. In the HT29 electropherogram peaks are absent below each arrow (except for some small ghost peaks at a few positions) indicating that HT29 is methylated at each cytosine in a CpG dinucleotide for these 150 base pairs. Good resolution is seen throughout both runs. The methylated DNA sequence did not suffer from poor resolution as did the methylated pUC 19 sequence as described in section 4.3.4. This could be because I used a T-termination sequencing reaction instead of an A-terminated reaction.

Before CD-PCR with this primer set is applied to the analysis of p16 in actual tumor samples it will need to be validated in the same manner as it was for the pUC 19 system in Chapter 5. Partial methylation of individual cytosines is not expected in cell lines that have undergone several passages. Clonal selection would ensure that all cells in sample have the same methylation status. Tumor samples are expected to contain cells with differing sites of methylation.

In this experiment I used a nested PCR, that is, I used an outer primer set and an inner primer set to enhance the specificity of the PCR reaction when working from a genomic DNA template. The total CD-PCR experiment involved 70 cycles of amplification; if even a small amount of bias was present at each cycle the error in the final measurements will be huge indeed. While I designed a nested PCR for this experiment it turns out that it wasn't necessary. I was able to get single band PCR

products from the deaminated DNA using only the outer primer set. Figure 6.2 is a photo of the agarose gel of PCR reactions done using only the inner primer set. The PCR reaction conditions used to obtain these bands were identical to those described in Section 6.2.4 except that 40 cycles of PCR were performed instead of 35. Not only should the use of a single PCR minimize bias but it also reduced reagent consumption and the time involved to complete an analysis.

While direct cycle sequencing of CD-PCR products is the fastest way to determine the methylation status of individual cytosines, p16 is a good example of a situation where cloning the CD-PCR products could provide valuable information. Schmidt et al. speculate that there are distinct subpopulations of cells in tumor samples, which are methylated at different sites[12]. Because cloning selects single CD-PCR product molecules for analysis, which arise from single chromosomes initially, cloning can determine if methylation at specific sites are related. For example, if two sites are determined to be partially methylated by direct sequencing of CD-PCR products, cloning CD-PCR products could distinguish between two cases. Case one: some cells have one cytosine methylated, some have the other cytosine methylated, and some cells have neither cytosine methylated. Case two: some cells have both cytosines methylated and some cells have neither cytosine methylated. Cloning the CD-PCR products is considerably more expensive and time consuming than direct sequencing, however the extra cost and effort may be justified in some cases.

6.4 Conclusion

The mechanism by which cytosine methylation is involved in gene silencing is unknown. The current data which correlates methylation of tumor suppressor genes to cancer is tantalizing indeed. Unfortunately, direct evidence that methylation is responsible for oncogenesis is lacking. Evidence of partial methylation of some cytosines and not at others in the p16 gene emphasizes the need for a methylation analysis system that can quantitatively determine the level of methylation for every cytosine in tumor suppressor genes and their promotors. The CD-PCR system that I have developed for p16 has the potential to generate quantitative data. This technique, combined with the measurement of p16 transcriptional activity, will hopefully begin to shed some light on these

unknowns.

Figure 6.1 T-termination sequencing electropherograms of CD-PCR products from A431 (top trace-solid line) and HT29 (bottom trace-dotted line) cell lines. Arrows indicate positions corresponding to cytosine in a CpG dinucleotide.

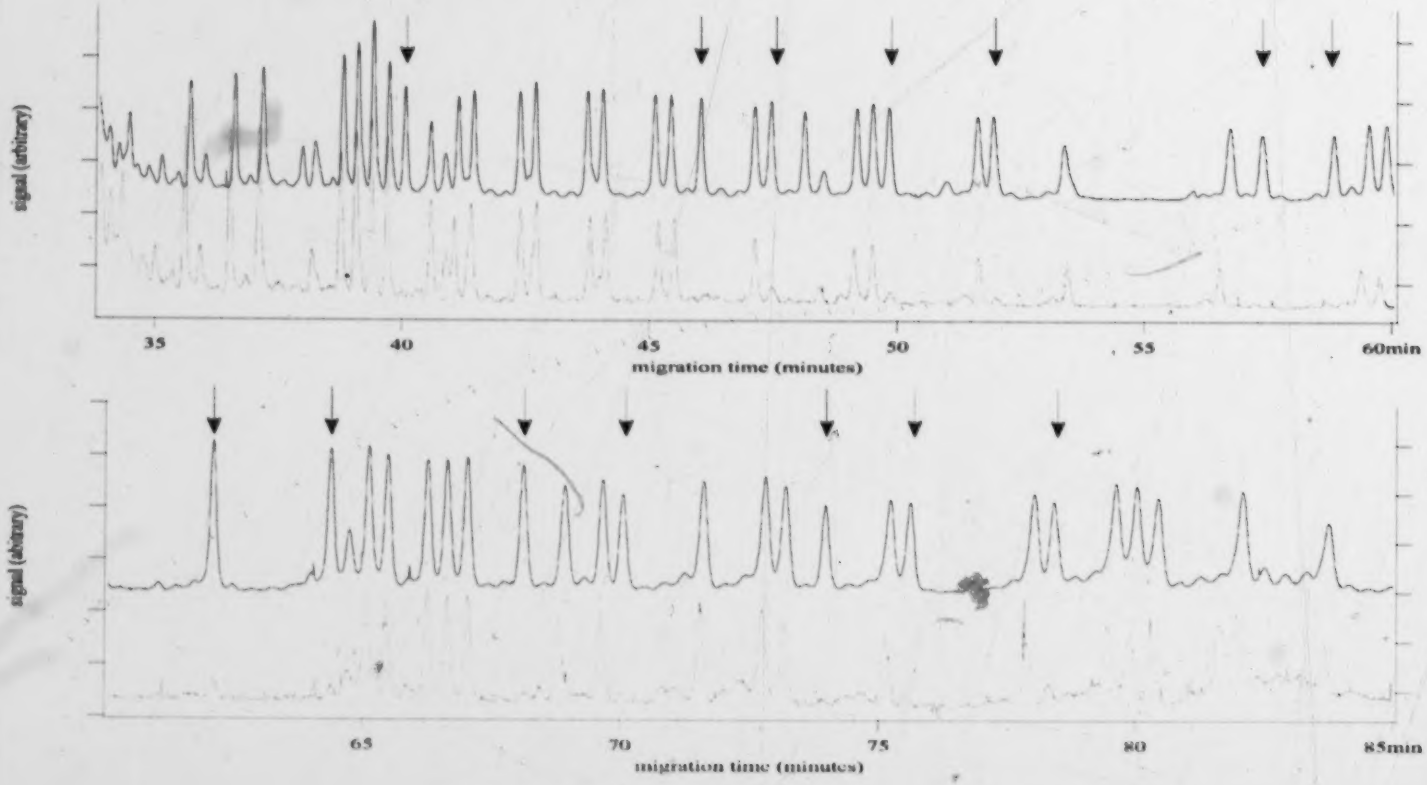


Figure 6.2 Photograph of agarose gel of PCR product from deaminated DNA using only the outer primer set. Lane 1 is the PCR product. Lane 2 is 1 kb ladder size-standard DNA.

6.5 Bibliography

1. Sakai, T.; Toguchida, J.; Ohtani, N.; Yandell, D. W.; Rapaport, J. M. and Dryja, T. P. *American Journal of Human Genetics* **48**, 880-888 (1991).
2. Greger, V.; Passarge, E.; Hopping, W.; Messmer, E. and Horsthemke, B. *Human Genetics* **83**, 155-158 (1898).
3. Herman, J. G.; Latif, F.; Weng, Y.; Lerman, M. I.; Zbar, B.; Liu, S.; Samid, D.; Duan, D. S.; Gnarra, J. R. and Linehan, W. M. *Proceedings of the National Academy of Sciences, USA* **91**, 9700-9704 (1994).
4. Merlo, A.; Herman, J. G.; Mao, L.; Lee, D. J.; Gabrielson, E.; Burger, P. C.; Balin, S. B. and Sidransky, D. *Nature Medicine* **1**, 686-692 (1995).
5. Gonzalez-Zulueta, M.; Bender, C. M.; Yang, A. S.; Nguyen, T.; Beart, R. W.; Tornout, J. M. B. and Jones, P. A. *Cancer Research* **55**, 4531-4535 (1995).
6. Herman, J. G.; Merlo, A.; Mao, L.; Lapidus, R. G.; Issa, J.-P. J.; Davidson, N. E.; Sidranski, D. and Baylin, S. B. *Cancer Research* **55**, 4525-4530 (1995).
7. Rideout, W. M.; Eversole-Cire, P.; Spruck, C. H.; Hustad, C. M.; Coetzee, G. A.; Gonzales, F. and Jones, P. A. *Molecular Cell Biology* **14**, 6143-6152 (1994).
8. Yoshiura, K.; Kanai, Y.; Ochiai, A.; Shimoyama, Y.; Sugimura, T. and Hirohashi, S. *Proceedings of the National Academy of Sciences, USA* **92**, 7416-7419 (1995).
9. Greunbaum, Y.; Cedar, H. and Razin, A. *Nature*, **295** (1982).
10. Serrano, M.; Hannon, G. J. and Beach, D. *Nature* **366**, 753-756 (1993).
11. Nobori. *Nature* **368**, 753-756 (1994).
12. Schmidt, E. E.; Ichimura, K.; Messerle, K. R.; Goike, H. M. and Collins, V. P. *British Journal of Cancer* **51**, 2-8 (1997).

CHAPTER 7: CONCLUSION

The data presented in Chapters 1 and 2 highlight the difficulties in identifying systematic sources of band broadening in capillary electrophoresis DNA sequencing. Not only do temperature fluctuations and capillary coating failure have similar empirical effects on the sequencing electropherograms, they also have similar effects on resolutions, theoretical plates, and peak widths. Worse yet, if other components have not yet been identified as a source of band broadening and are randomly changed, it is impossible to obtain reproducible results. It is necessary to consider all components of the sequencing system as potential sources of band broadening.

However, the work in Chapters 1 and 2 demonstrates that it is possible to systematically study DNA sequencing by capillary electrophoresis. In their paper titled "Why can we not sequence thousands of DNA bases on a polyacrylamide gel" Slater and Drouin examine the fundamental limits of DNA sequencing in polyacrylamide slab gels[1]. An equally thorough examination of the fundamental limits of DNA sequencing by CE and entangled polymer solutions needs to be developed. With most of the reproducibility problems that have plagued CE sequencing now under control, it is possible to experimentally test the validity of the electrophoresis theories discussed in Chapter 1 as they apply to CE. In particular, the effects of electric field strength, separation temperature, capillary length, polymer viscosity and molecular weight, buffer composition, and choice of denaturant have not been systematically investigated. Studies of these parameters would be appropriate for future work.

As this thesis has shown, not all limits to CE-DNA sequencing performance are fundamental in nature. I have examined the effect of temperature stability and coating stability on CE-DNA sequencing. Other parameters, such as physical damage to the capillaries, also need to be examined. Perhaps most importantly, the effects of sample injection and sample load on the quality of sequencing separations needs to be addressed. I believe that there is a limit to the absolute mass of DNA that can be loaded onto a capillary filled with sieving matrix without. This limit would place a crucial constraint on the sensitivity required for detection systems in multicapillary instrumentation. Design constraints for detector sensitivity must be determined if attempts at commercializing CE-DNA sequencing instrumentation are likely to succeed.

While different in its approach, the developing new chemistries for DNA methylation analysis has much the same goal as the developing CE-DNA sequencing technologies. Both studies are dedicated to developing better tools to learn more about our genetic makeup-wherein is hidden many of the mysteries of who and what we are.

DNA methylation researchers do not yet have an equivalent of Sanger sequencing chemistry at their disposal. CD-PCR chemistry, while promising high sensitivity analysis of all cytosines, is not likely to develop into the high throughput technology that the Sanger chemistry is to the DNA sequencing world. I believe that better technology must be developed, but will require more than just a great deal of work. It requires new ideas. One possible advance might be methylation-preserving PCR. Perhaps a thermostable maintenance methylase can convert hemi-methylated sites to fully methylated sites following each cycle of replication in a PCR. This would allow biological researchers to apply the powerful PCR chemistry directly to methylation analysis.

Despite its limitations, CD-PCR chemistry is already a valuable tool for biological research. CD-PCR and the improvements that I introduced in Chapters 4, 5 and 6 are being used by other members of Dovichi's group. The CD-PCR chemistry described in Chapter 6 is now being used to examine changes in methylation profiles of the p16 gene in cell lines due to varying dietary folate levels. CD-PCR chemistry can, in principle, be applied to any DNA sequence. Work is already underway to use CD-PCR chemistry to study methylation of the BRCA-2 tumor suppressor gene.

Better tools for methylation analysis will lead to more interest in DNA methylation biology, which in turn will push the search for better tools to study methylation. As the Human Genome Project progresses most human genes will be discovered. Some researchers will wish to determine how methylation is involved in the regulation of their favorite gene. As long as the interest in DNA methylation continues to grow, and it will, there will be a need for better bio-analytical tools.

7.2 Bibliography

1. Slater, G. W. and Drouin, G. *Electrophoresis* **13**, 574-582 (1992).